PALMs: Using Multi Construct-Grounded Rationales for Modeling Population Preferences in LLMs

Priyanka Dey[1,2] **Brihi Joshi**[1] **Preyashi Poddar**[1] **Jieyu Zhao**[1] **Emilio Ferrara**[1,2]
[1]University of Southern California [2]Information Sciences Institute
{deyp}@usc.edu

Abstract

Large language models are being extensively used to simulate individual user behavior, yet faithfully representing a *population* requires capturing the systematic variation in values, beliefs, and cultural norms that distinguish one group from another. We introduce *Population Aligned Language Models* (*PALMs*), a suite of models each aligned to specific populations, covering five countries: USA, India, Brazil, France and Italy. *PALMs* are created by synthesizing rationales grounded in psychological and cultural constructs and using these as latent supervision during preference tuning for population-specific alignment. Evaluated across four dimensions: personality, values and beliefs, cultural norms, and morality, *PALMs* consistently outperform baselines, including culture-specialized models, achieving an average of 8.59% relative improvement over the best baseline across all five populations. Notably, construct-grounded rationales outperform both demographic prompting and survey-based fine-tuning, suggesting that grounding preference learning in psychology and culture provides a richer inductive signal than surface-level response distributions. We further demonstrate strong generalization to downstream applications without task-specific supervision: outperforming best baselines by 5.19% in personalized reward modeling, 6.34% in population simulation, and showing strong transfer to social reasoning tasks. Datasets and code are available at: https://github.com/limenlp/PALMs.

1 Introduction

While LLMs serve billions of users worldwide, a core challenge in alignment remains: moving beyond a single, aggregate notion of human preference toward models that reflect the plurality of values and preferences across cultures (Sorensen et al., 2024; Santurkar et al., 2023; Aroyo et al., 2023; Prabhakaran et al., 2022; Durmus et al., 2023). User simulation — modeling how specific users or populations respond to language — has emerged as a key testbed, enabling applications such as personalized reward modeling, opinion prediction, and social reasoning (Argyle et al., 2023; Wright et al., 2024; Park et al., 2022; Krsteski et al., 2025). Yet existing frameworks condition models on shallow demographic attributes or short persona descriptions that fail to capture the values, beliefs, and cultural norms distinguishing one population from another (Santurkar et al., 2023; AlKhamissi et al., 2024; Dominguez-Olmedo et al., 2024; Sun et al., 2025; Gupta et al., 2024). Fine-tuning directly on behavioral or survey data does not reliably solve this either, as surface-level optimization fails to preserve the richer compositional structure of human preferences (Neumann et al., 2026; Li et al., 2024a; Suh et al., 2025; Meister et al., 2025).

Human preferences are not flat outputs of demographics (Alipour et al., 2025; Torelli and Kaikati, 2009; Hwang et al., 2023). They emerge from the interaction of multiple psychological and cultural factors: personality traits (John, 1999), cultural dimensions (Hofstede, 2001), moral foundations (Graham et al., 2013), personal values (Schwartz, 1992), and world beliefs (Clifton, 2020). While many such constructs exist, we focus on these five as they span individual-level psychology and group-level cultural variation, providing complementary lenses for reasoning about population-level preferences. Existing approaches either prompt models with demographic data or fine-tune on observed responses, neither of which forces LLMs to represent the underlying constructs that jointly shape human judgment.

In this work, *population* refers to country-level human groups whose preferences reflect shared cultural, social, and behavioral patterns.[1] We in-

[1]We acknowledge that meaningful variation exists within countries across regional, linguistic, and socioeconomic lines (Orlikowski et al., 2023); we adopt country-level abstraction

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 20475–20517
October 24-29, 2026 ©2026 Association for Computational Linguistics

troduce ***Population Aligned Language Models*** (***PALMs***), a suite of LLMs each aligned to a specific population across five countries: USA, India, Brazil, France, and Italy. *PALMs* are trained using *multi-construct rationales*: synthetic rationales grounded in five psychological and cultural constructs — personality traits (OCEAN; Goldberg 1993), cultural dimensions (Hofstede; Hofstede 2001), human values (Schwartz; Schwartz 1992), moral foundations (Graham et al., 2013), and world beliefs (Clifton, 2020) — as latent supervision during preference optimization via DPO (Rafailov et al., 2023), with rationale tokens masked from the loss. At inference, *PALMs* generate rationales before producing outputs, requiring no external construct annotation. Figure 1 shows the training and inference pipeline.

We evaluate *PALMs* across four dimensions of population alignment — personality, values and beliefs, cultural norms, and moral values — by sampling users conditioned on demographic backgrounds drawn from each target population and aggregating their responses against ground-truth human distributions. *PALMs* consistently outperform demographic prompting baselines and culture-specialized models across all five populations, achieving an average 8.59% relative improvement over the best baseline. Notably, fine-tuning on survey data can underperform zero-shot baselines, suggesting that surface-level optimization may collapse rather than enrich population representations, whereas *PALMs* avoid this by grounding preference reasoning in latent constructs. Beyond population alignment, *PALMs* also show strong performance in three downstream applications without task-specific supervision: personalized reward modeling (5.19% improvement), population simulation (6.34% improvement), and social reasoning.

Our contributions are as follows: (1) We introduce ***Population Aligned Language Models*** (***PALMs***), a suite of population-aligned LLMs built on multi-construct rationales grounded in established psychological, cultural, and values theory. (2) We demonstrate that construct-grounded rationales outperform both demographic prompting and behavioral fine-tuning across five countries and four alignment dimensions. (3) We show that *PALMs* generalize without task-specific supervision to several downstream tasks, establishing multi-construct rationales as a principled foundation for modeling populations.

as a practical and widely-used unit of analysis (Li et al., 2024a; Khanuja et al., 2024; Liu et al., 2025b).

2 Related Work

Persona-based user modeling. A central line of work in LLM personalization conditions model behavior on user personas — descriptions comprising demographic attributes, prior judgments, or interaction history (Mazaré et al., 2018; Madotto et al., 2019). Zero-shot demographic prompting is conceptually appealing but produces inconsistent results: models tend to over-represent majority viewpoints and reflect known training biases (Santurkar et al., 2023; Hu et al., 2025b), and sociodemographic prompting does not consistently improve — and in some cases worsens — alignment with specific subpopulations (Sun et al., 2025). More recent approaches incorporate prior user judgments into the prompt (Hwang et al., 2023). Joshi et al. (2025), Koncel-Kedziorski et al. (2025) augment personas with psychologically scaffolded rationales to explain user judgments at inference time, and Dey et al. (2026) generate synthetic profile-grounded preference pairs using cultural and psychological frameworks to improve personalized content generation. A related line of work builds personalized *reward models* directly: PAL (Chen et al., 2025) learns modular, user-specific reward heads for few-shot preference localization, PersRM-R1 (Li et al., 2025) incorporates reinforcement learning to improve personalized reward modeling, and Ryan et al. (2025) induce persona-guided prompts to condition reward models on inferred user traits. Our work differs from all of these: rather than enriching inference-time inputs or specializing the reward model architecture, we use construct-grounded reasoning as *training-time* latent supervision, shaping the model's internal representations through masked scratchpads during preference optimization.

Cultural and population alignment. A growing body of work studies how well LLMs represent diverse human populations, motivated by the broader case for pluralistic alignment (Sorensen et al., 2024). (Santurkar et al., 2023) and (Argyle et al., 2023) show that LLMs can simulate survey responses but tend to over-represent majority viewpoints. Culture-specialized models such as CultureLLM (Li et al., 2024a) and CulturePark (Li et al., 2024b) train directly on large-scale cultural survey data to improve cross-cultural representation. (Meister et al., 2025) and (Dey et al., 2025)

propose frameworks for evaluating LLM alignment with human personality and opinion distributions. Our work builds on this evaluation paradigm but identifies a failure mode these approaches share: optimizing directly on behavioral outputs narrows representational diversity rather than enriching it, a pattern we observe consistently across our experiments.

Reasoning as training supervision. Chain-of-thought prompting (Wei et al., 2022) and its variants have shown that eliciting intermediate reasoning steps improves LLM performance across a range of tasks. STaR (Zelikman et al., 2022) and related methods use model-generated rationales as self-improvement signals during training. Scratchpad-based approaches (Nye et al., 2021) similarly use intermediate computation as latent structure. In preference learning, Lightman et al. (2024) show that process-level supervision outperforms outcome-level supervision for mathematical reasoning. Our approach is in this spirit but targets a different goal: we use reasoning not to improve task accuracy, but to inject structured psychological knowledge into preference representations, masking it from the optimization objective so it functions purely as inductive bias.

Psychological constructs in language models. Humans routinely use everyday explanations to rationalize each other's actions—a practice known as folk psychology (Malle, 2006; Churchland, 2013). Such explanations are prevalent enough in text corpora that LLMs can generate plausible psychological rationales zero-shot (Binz and Schulz, 2023), making synthetic construct-grounded reasoning chains a tractable source of structured supervision. Prior work has applied individual constructs to NLP tasks: Big Five traits have been used to model user behavior (Goldberg, 1993), Schwartz values to study opinion formation (Schwartz, 1992; Kiesel et al., 2022), and moral foundations to analyze political discourse (Haidt and Joseph, 2004). *Population Aligned Language Models* unifies these into a single scaffold space, using them jointly as lenses for reasoning about preference decisions.

3 *Population Aligned Language Models*

3.1 Problem Formulation

Consider a user from a country (e.g., USA, India), described by an individual demographic persona p (e.g. their ethnicity, age, gender, education level, and political leaning). Given a query q paired with two candidate responses a^+ and a^-, our goal is to predict the user's preferred response a^*.

Standard preference learning optimizes a direct mapping $(p, q) \rightarrow a^*$, conditioning only on surface-level demographic descriptors. *Population Aligned Language Models* instead introduces an intermediate multi-construct rationale r, decomposing each decision across five psychological and cultural constructs (§3.2) before predicting the preferred response: $(p, q) \rightarrow r \rightarrow a^*$. The rationale r is supplied as latent supervision during training and masked from the loss, so the model internalizes construct-grounded rationale without learning to generate the rationale token-by-token. At inference time, the model generates $r*$ before producing a^*, requiring no external annotation.

Generating this rationale r allows the model to capture the psychological and cultural mechanisms behind a preference, rather than fitting surface-level correlations between demographic descriptors and outputs. A direct $(p, q) \rightarrow a^*$ mapping does not provide a mechanism to represent *why* a person with a given background prefers one response over another. By generating r first, the model reasons through an individual's personality traits, values, moral intuitions, and world beliefs, situating them within the broader cultural tendencies that shape how those traits are expressed (Sagiv and Schwartz, 2022; Vauclair et al., 2015).

3.2 Construct Space

We ground our rationales in five families of established social science theory that together span individual-level psychology and group-level cultural variation. These construct families act as interpretive lenses through which the model reasons about each preference decision (See Table 1).

Figure 2 illustrates an example synthetic rationale generated from these five families for a user's preference instance.

3.3 Construct-Grounded Rationale Generation

For each training instance (p, q, a^+, a^-), we prompt an LLM f_ϕ to generate construct-grounded rationales that explain the observed preference as a function of the user's demographic profile. These post-hoc rationalizations provide a justification for *why* a user with profile p might plausibly favor a^+ over a^- through each construct lens—not inferred ground truths about the user's

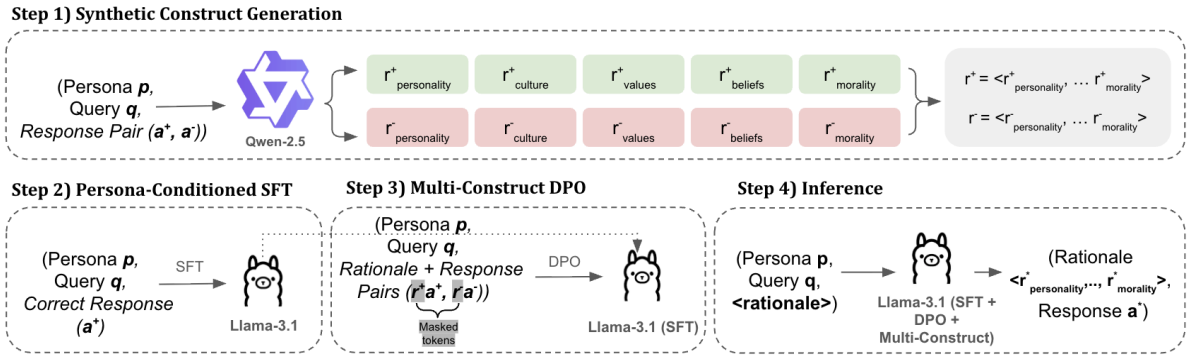

Figure 1: Overview of the *Population Aligned Language Models* training and inference pipeline. We generate multi-construct post-hoc rationales grounded in $\mathcal{C}$ (see Table 1), then run SFT on persona-conditioned preference instances. The SFT model is further optimized with DPO on rationale-augmented preference pairs, with rationale tokens masked from the loss and used only as structured context. At inference, given a persona, query, and `<rationale>` token, the model generates a construct-grounded rationale followed by a preference prediction.

Construct Family	Description
Personality (OCEAN) (Goldberg, 1993)	Individual-level traits (e.g., openness, conscientiousness) that shape communication and decision style.
Cultural dimensions (Hofstede) (Hofstede, 2001)	Group-level tendencies (e.g., individualism, power distance) that govern how people relate to authority, community, and social expectations.
Human values (Schwartz) (Schwartz, 1992)	Individual-level motivational priorities (e.g., conservation, self-transcendence) that drive trade-offs between competing choices.
Primal world beliefs (Clifton, 2020)	Individual-level assumptions about the world (e.g., safe vs. dangerous) that determine situational interpretation.
Moral foundations (Haidt and Joseph, 2004)	Individual-level moral intuitions (e.g., fairness, authority) that govern ethical evaluation.

Table 1: Construct families spanning the latent space $\mathcal{C}$.

inner states, but plausible explanations consistent with the observed demographics and choice. For each construct $C_k \in \mathcal{C}$ and each candidate response $a \in \{a^+, a^-\}$, we condition f_ϕ on the demographics, query, candidate response, and preference label $\ell \in \{\text{chosen}, \text{rejected}\}$ to produce a construct-specific reasoning chain $r^{(k)}$, each consisting of 4–6 steps. The full rationale concatenates all construct-specific chains: $r = \langle r^{\text{PERS}}, r^{\text{CULT}}, r^{\text{VAL}}, r^{\text{MOR}}, r^{\text{BEL}} \rangle$.

3.4 Training

We train *Population Aligned Language Models* in two stages, with a separate model per country (see Fig 1).

Stage 1: Supervised Fine-Tuning. We fine-tune `Llama-3.1-8B-Instruct` on preference instances (p, q, a^+) with a standard causal LM objective to initialize the model on persona-conditioned preference prediction before preference optimization.

Stage 2: Preference Optimization. We then augment each preference pair with the rationales from §3.3, where r^+ rationalizes a^+ as the chosen response and r^- rationalizes a^- as the rejected one: $c^\pm = \langle$`<rationale>` $r^\pm$ `</rationale>`, $a^\pm \rangle$. We optimize c^+ and c^- using Direct Preference Optimization (DPO; Rafailov et al. 2023), masking all rationale tokens from the loss and compute loss over over a^+ and a^-. As multiple rationales are equally plausible for any given instance, optimizing directly over them would overfit the model to arbitrary explanations. Masking ensures the rationales instead shape preference learning as structured context rather than explicit prediction targets.

Inference. During inference, *PALMs* are prompted with the `<rationale>` token and generate a construct-grounded rationale before producing a final preference decision. Only the final output is used for evaluation.

4 Experiments

Training Details. We train *Population Aligned Language Models* on the Community Alignment dataset (Zhang et al., 2026), a large-scale preference dataset designed capturing population-grounded human judgments. Each sample consists of a user persona with associated demographic attributes, a query, and a pair of candidate responses with a preferred choice. We augment each preference pair with synthetic construct-grounded rationales generated by `Qwen2.5-7B-Instruct` as described in §3.3. We train five separate models, one per target population: USA, India, Brazil, France, and Italy (see App. B for dataset statis-

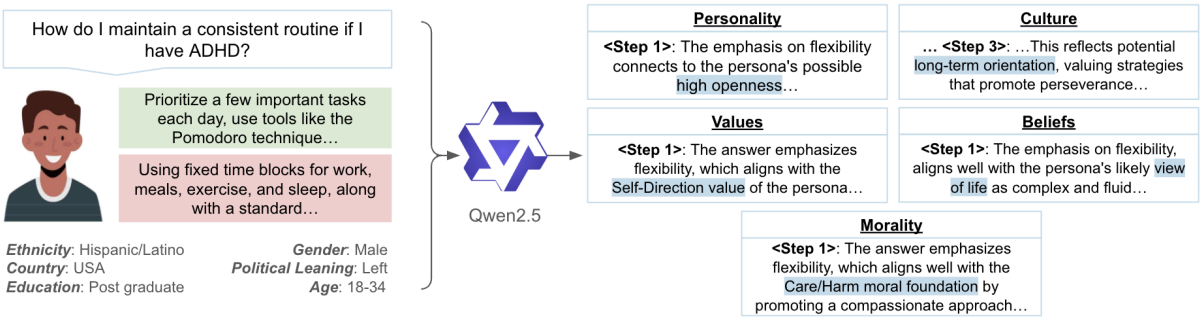

Figure 2: During training, we prompt `Qwen2.5-7B-Instruct` to produce synthetic construct-grounded rationales for each preference instance. Each rationale decomposes the preference decision across five psychological and cultural construct families, providing structured label-conditioned supervision during DPO optimization.

tics and training details). We train five separate models, one per target population: USA, India, Brazil, France, and Italy (see App. B for dataset statistics and training details), initialized from `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) and trained following the two-step pipeline in §3.4.

In-Domain Evaluation. We evaluate population alignment across four dimensions: *personality, culture, values and beliefs*, and *morality*. For each dimension, we select an established survey instrument and measure alignment between model response distributions (via sampling users by conditioning on demographic backgrounds) and ground-truth human distributions from the target population.

For *personality*, we follow Dey et al. (2025) and Meister et al. (2025). We prompt each *PALM* to produce response distributions over personality items and measure alignment via Wasserstein distance across the five OCEAN traits (Goldberg, 1999; Johnson, 2014). For *culture*, we use open-ended cultural norm questions from CultureBank (Shi et al., 2024) and measure alignment via entailment accuracy using GPT-4o (Hurst et al., 2024) as an LLM judge (Zheng et al., 2023). For *values and beliefs* and *morality*, we use opinion questions from the Pew Global Attitudes Survey (Pew Research Center, 2023) and the World Values Survey (Haerpfer et al., 2022) respectively, measuring population-level alignment via L_1 distance between model and human response distributions over unordered categorical options. We normalize all metrics to $[0, 1]$ and apply a $1 -$ score transformation to lower-is-better metrics (Wasserstein, L_1) before averaging, so that higher scores consistently reflect better alignment across all four dimensions.

Out-of-Domain Evaluation. To assess whether construct-grounded representations generalize beyond the training objective, we evaluate *Population Aligned Language Models* on three downstream applications where modeling diverse human perspectives is critical but existing approaches remain limited: (1) *personalized reward modeling*, (2) *population simulation*, and (3) *social intelligence*.

For *personalized reward modeling*, general-purpose reward models are not tailored to specific user populations, yet collecting group-specific preference data is often infeasible. We evaluate on *PersonalizedRewardBench* (Ma et al., 2026), a personalized reward benchmark spanning datasets across varied topics like lifestyle, culture, and entertainment, and *PRISM* (Kirk et al., 2024), a globally diverse preference dataset focused on values and opinions. We benchmark *PALMs* against Skywork-Reward-V2 (Liu et al., 2025a), an 8B model fine-tuned specifically for reward modelling and a top performer on RewardBench (Lambert et al., 2025) and PAL (Chen et al., 2025), a 350M-parameter reward model designed for pluralistic, personalized preference modeling, as strong baselines.

For *population simulation*, LLMs are increasingly used to simulate human behavior, but struggle to capture variation across real human populations. We evaluate PALMs on SIMBENCH (Hu et al., 2025a), a large-scale benchmark spanning opinion, personality, moral reasoning, humor, and decision-making. We report SimBench's normalized score derived from Total Variation Distance, where 100 indicates perfect alignment, 0 equals random guessing, and negative scores indicate worse-than-random performance.

For *social intelligence*, interpretations of social situations vary systematically across cultural norms, values, and personality — yet existing

Model	Country	Overall ↑	Personality ↑	Culture ↑	Values ↑	Morality ↑
Baselines						
Llama-3.1-8B-Instruct	USA	$0.522_{\pm 0.014}$	$0.541_{\pm 0.007}$	$\underline{0.527}_{\pm 0.029}$	$0.519_{\pm 0.046}$	$0.502_{\pm 0.004}$
	India	$0.490_{\pm 0.033}$	$0.487_{\pm 0.010}$	$\underline{0.475}_{\pm 0.123}$	$0.524_{\pm 0.044}$	$0.475_{\pm 0.004}$
	Brazil	$0.468_{\pm 0.030}$	$0.431_{\pm 0.017}$	$\underline{0.436}_{\pm 0.108}$	$0.511_{\pm 0.044}$	$0.492_{\pm 0.004}$
	France	$0.491_{\pm 0.018}$	$0.508_{\pm 0.014}$	$0.495_{\pm 0.054}$	$0.532_{\pm 0.044}$	$0.428_{\pm 0.004}$
	Italy	$0.487_{\pm 0.016}$	$0.524_{\pm 0.020}$	$0.455_{\pm 0.046}$	$0.516_{\pm 0.040}$	$0.452_{\pm 0.004}$
Gemini-2.5-Flash	USA	$0.642_{\pm 0.025}$	$0.644_{\pm 0.006}$	$0.507_{\pm 0.029}$	$0.630_{\pm 0.094}$	$\underline{0.787}_{\pm 0.004}$
	India	$0.646_{\pm 0.032}$	$\underline{0.719}_{\pm 0.008}$	$0.390_{\pm 0.120}$	$\underline{0.760}_{\pm 0.046}$	$0.717_{\pm 0.004}$
	Brazil	$0.592_{\pm 0.030}$	$0.550_{\pm 0.016}$	$0.397_{\pm 0.106}$	$0.694_{\pm 0.054}$	$\underline{\mathbf{0.726}}_{\pm 0.004}$
	France	$0.634_{\pm 0.019}$	$0.640_{\pm 0.013}$	$0.480_{\pm 0.054}$	$\underline{\mathbf{0.735}}_{\pm 0.051}$	$0.681_{\pm 0.004}$
	Italy	$0.643_{\pm 0.017}$	$\underline{0.709}_{\pm 0.015}$	$0.401_{\pm 0.045}$	$0.710_{\pm 0.048}$	$0.754_{\pm 0.004}$
CultureLLM	USA	$0.480_{\pm 0.012}$	$0.499_{\pm 0.006}$	$0.513_{\pm 0.029}$	$0.479_{\pm 0.040}$	$0.430_{\pm 0.004}$
	India	$0.452_{\pm 0.031}$	$0.568_{\pm 0.009}$	$0.373_{\pm 0.119}$	$0.454_{\pm 0.036}$	$0.413_{\pm 0.004}$
	Brazil	$0.415_{\pm 0.029}$	$0.427_{\pm 0.017}$	$0.359_{\pm 0.104}$	$0.474_{\pm 0.044}$	$0.400_{\pm 0.004}$
	France	$0.457_{\pm 0.016}$	$0.478_{\pm 0.015}$	$\underline{0.517}_{\pm 0.053}$	$0.421_{\pm 0.036}$	$0.413_{\pm 0.004}$
	Italy	$0.436_{\pm 0.015}$	$0.506_{\pm 0.019}$	$0.414_{\pm 0.045}$	$0.420_{\pm 0.035}$	$0.405_{\pm 0.004}$
Qwen2.5-72B-Instruct	USA	$\underline{0.675}_{\pm 0.018}$	$\underline{0.690}_{\pm 0.010}$	$0.500_{\pm 0.036}$	$\underline{\mathbf{0.730}}_{\pm 0.039}$	$0.780_{\pm 0.005}$
	India	$\underline{0.663}_{\pm 0.026}$	$0.680_{\pm 0.011}$	$\underline{0.475}_{\pm 0.116}$	$0.745_{\pm 0.041}$	$\underline{\mathbf{0.750}}_{\pm 0.005}$
	Brazil	$\underline{0.618}_{\pm 0.024}$	$\underline{\mathbf{0.640}}_{\pm 0.015}$	$0.430_{\pm 0.103}$	$\underline{\mathbf{0.700}}_{\pm 0.043}$	$0.700_{\pm 0.005}$
	France	$\underline{0.646}_{\pm 0.019}$	$\underline{\mathbf{0.660}}_{\pm 0.012}$	$0.485_{\pm 0.057}$	$0.725_{\pm 0.040}$	$\mathbf{0.715}_{\pm 0.005}$
	Italy	$\underline{0.660}_{\pm 0.021}$	$0.700_{\pm 0.014}$	$\underline{0.460}_{\pm 0.049}$	$\underline{\mathbf{0.710}}_{\pm 0.042}$	$\mathbf{0.780}_{\pm 0.005}$
PALMs						
SFT	USA	$0.655_{\pm 0.009}$	$0.590_{\pm 0.010}$	$0.650_{\pm 0.022}$	$0.620_{\pm 0.024}$	$0.760_{\pm 0.008}$
	India	$0.620_{\pm 0.030}$	$0.560_{\pm 0.012}$	$0.550_{\pm 0.115}$	$0.690_{\pm 0.032}$	$0.680_{\pm 0.007}$
	Brazil	$0.558_{\pm 0.026}$	$0.490_{\pm 0.014}$	$0.410_{\pm 0.098}$	$0.660_{\pm 0.030}$	$0.670_{\pm 0.008}$
	France	$0.580_{\pm 0.014}$	$0.590_{\pm 0.011}$	$0.460_{\pm 0.048}$	$0.660_{\pm 0.028}$	$0.610_{\pm 0.009}$
	Italy	$0.563_{\pm 0.014}$	$0.490_{\pm 0.016}$	$0.420_{\pm 0.042}$	$0.650_{\pm 0.031}$	$0.690_{\pm 0.007}$
SFT + DPO	USA	$0.683_{\pm 0.008}$	$0.610_{\pm 0.009}$	$0.700_{\pm 0.020}$	$0.640_{\pm 0.022}$	$0.780_{\pm 0.007}$
	India	$0.653_{\pm 0.029}$	$0.580_{\pm 0.011}$	$0.600_{\pm 0.112}$	$0.720_{\pm 0.030}$	$0.710_{\pm 0.006}$
	Brazil	$0.588_{\pm 0.025}$	$0.520_{\pm 0.013}$	$0.440_{\pm 0.095}$	$0.690_{\pm 0.028}$	$0.700_{\pm 0.007}$
	France	$0.613_{\pm 0.013}$	$0.620_{\pm 0.010}$	$0.500_{\pm 0.045}$	$0.690_{\pm 0.026}$	$0.640_{\pm 0.008}$
	Italy	$0.590_{\pm 0.013}$	$0.510_{\pm 0.015}$	$0.450_{\pm 0.040}$	$0.680_{\pm 0.029}$	$0.720_{\pm 0.006}$
SFT + DPO + Generic	USA	$0.699_{\pm 0.008}$	$0.622_{\pm 0.008}$	$0.729_{\pm 0.019}$	$0.650_{\pm 0.021}$	$0.795_{\pm 0.006}$
	India	$0.676_{\pm 0.028}$	$0.590_{\pm 0.010}$	$0.644_{\pm 0.108}$	$0.740_{\pm 0.029}$	$\mathbf{0.730}_{\pm 0.005}$
	Brazil	$0.600_{\pm 0.024}$	$0.530_{\pm 0.012}$	$0.455_{\pm 0.092}$	$0.700_{\pm 0.026}$	$\mathbf{0.715}_{\pm 0.006}$
	France	$0.632_{\pm 0.012}$	$0.638_{\pm 0.009}$	$0.520_{\pm 0.042}$	$0.715_{\pm 0.024}$	$0.656_{\pm 0.007}$
	Italy	$0.610_{\pm 0.012}$	$0.526_{\pm 0.014}$	$0.470_{\pm 0.038}$	$0.700_{\pm 0.027}$	$0.745_{\pm 0.005}$
SFT + DPO + Multi-Construct	USA	$\mathbf{0.726}_{\pm 0.008}$	$\mathbf{0.656}_{\pm 0.006}$	$\mathbf{0.772}_{\pm 0.024}$	$0.672_{\pm 0.018}$	$\mathbf{0.802}_{\pm 0.004}$
	India	$\mathbf{0.738}_{\pm 0.026}$	$\mathbf{0.728}_{\pm 0.008}$	$\mathbf{0.714}_{\pm 0.102}$	$\mathbf{0.788}_{\pm 0.025}$	$0.722_{\pm 0.004}$
	Brazil	$\mathbf{0.618}_{\pm 0.023}$	$0.565_{\pm 0.012}$	$\mathbf{0.490}_{\pm 0.088}$	$\mathbf{0.708}_{\pm 0.024}$	$0.708_{\pm 0.004}$
	France	$\mathbf{0.647}_{\pm 0.011}$	$0.650_{\pm 0.009}$	$\mathbf{0.558}_{\pm 0.038}$	$0.730_{\pm 0.022}$	$0.650_{\pm 0.004}$
	Italy	$\mathbf{0.703}_{\pm 0.011}$	$\mathbf{0.720}_{\pm 0.012}$	$\mathbf{0.608}_{\pm 0.035}$	$\mathbf{0.715}_{\pm 0.024}$	$0.771_{\pm 0.004}$

Table 2: In-domain alignment results across five target populations. Higher scores are better; subscripts denote bootstrapped 95% confidence intervals. **Bold** indicates the best result per country–dimension; underlines indicate the best baseline. *Population Aligned Language Models* consistently achieves the highest overall alignment, with each training stage contributing incrementally. `CultureLLM` frequently underperforms zero-shot baselines, suggesting that survey-based fine-tuning does not reliably improve population alignment.

benchmarks largely treat social reasoning as uniform across individuals. We conduct a case study using *PALM*-USA on SOCIAL IQA (Sap et al., 2019), a benchmark for reasoning about human intent, emotion, and interpersonal expectations.[2]

Baselines. All models receive demographics per user: ethnicity, age group, gender, education, and political leaning, as provided by Community Alignment. We use three open- and closed-source zero-shot baselines: Llama-3.1-8B-Instruct, Qwen2.5-72B-Instruct, and Gemini-2.5-Flash. We also include CultureLLM (Li et al., 2024a), a culture-specialized model trained on large-scale human survey data.

To isolate the contribution of construct-grounded rationales, we compare four configurations: *SFT*, fine-tuned on the community dataset alone; *SFT+DPO*, which adds preference optimization without rationales; *SFT+DPO+Generic*, which incorporates post-hoc rationales *without* construct grounding; and *Population Aligned Language Models*, the full model. For out-of-domain evaluation,

[2]We restrict this evaluation to the USA population as SOCIAL IQA does not provide cross-cultural annotations for our other target populations, making a controlled single-population analysis the most meaningful comparison.

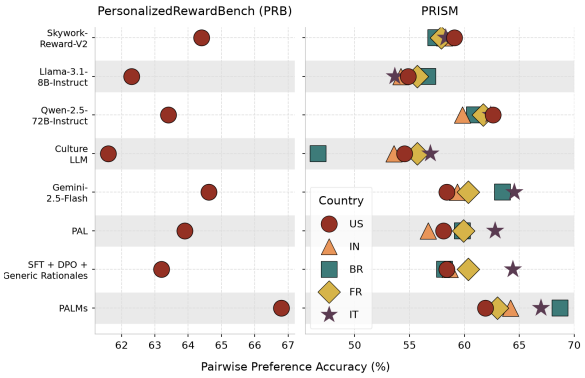

Figure 3: Pairwise preference prediction accuracy (%) on *PersonalizedRewardBench* (PRB) and *PRISM*, aggregated by country. *PALMs* achieves the strongest performance on PRB and on four of the five PRISM country splits.

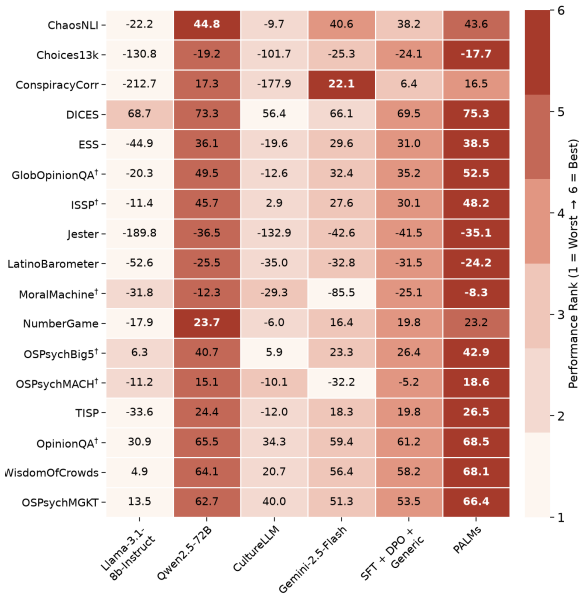

Figure 4: Population simulation results on SIMBENCH, macro-averaged over countries. Color indicates the within-row performance rank (1 = worst and 6 = best). † denotes tasks aligned with construct families (personality, culture, values, and morality). *PALMs* outperforms baselines on most opinion, personality, value-aligned, and behavioral simulation tasks.

we restrict comparison to *Population Aligned Language Models* and *SFT+DPO+Generic*.

5 Results

5.1 In-Domain Population Alignment

Table 2 presents results on in-domain population alignment across all five target countries.

***PALMs* achieves the strongest alignment across populations and dimensions.** *PALMs* consistently achieves the highest overall alignment across all countries, outperforming both baselines and intermediate training configurations. Within the constructs used, *culture* shows the largest relative improvement over baselines, while *personality* shows the most consistent gains across the training progression. *Morality* is the most variable construct, where intermediate models occasionally match *PALMs*, suggesting that moral alignment may require construct scaffolds beyond those captured by our current taxonomy.[3]

Construct-grounded rationales provide the strongest alignment signal at each training stage. Moving from *SFT* to *SFT+DPO* yields consistent but modest gains, showing that contrastive preference optimization sharpens population-specific alignment beyond supervised training alone. Adding generic rationales (*SFT+DPO+Generic*) produces a further improvement across most dimensions, showing that structured intermediate reasoning helps even without psychological grounding. The gap between *SFT+DPO+Generic* and *PALMs* isolates what construct grounding specifically contributes: for the USA, culture alignment improves from 0.729 to 0.772 and personality alignment improves from 0.622 to 0.656. Similar patterns appear across all five populations, suggesting that the construct families provide a decomposition vocabulary that generic reasoning alone cannot replicate. Additionally, a single-population leave-one-construct-out ablation (Appendix G) confirms that each construct contribute complementary signal.

Fitting directly to user behavioral data does not always improve population alignment. Models optimized on user-level survey responses without construct grounding show diminishing returns and, in some cases, underperform zero-shot baselines, particularly on culture and morality. Directly fitting observed response distributions does not reliably capture the diversity of human preferences across populations. Rationalizing through latent psychological and cultural constructs instead appears to provide a more robust path to alignment across diverse populations.

[3]Unlike personality, values, and cultural norms, moral judgments are highly sensitive to situational framing (Haidt and Joseph, 2004), such that demographic and construct-level signals may be insufficient to capture within-population moral variation.

5.2 *Application 1*: Personalized Reward Modeling

Construct-grounded rationales transfer to personalized reward modeling without any task-specific supervision. As shown in Fig. 3, *PALMs* achieves the strongest performance on PRB and on four of the five PRISM country splits. On PRB, *PALMs* reaches 66.80% against Skywork's 64.40%, an improvement of 2.40 percentage points. We further observe consistent gains across PRB topic categories (Table 7), with the largest improvements on society-related topics spanning religious beliefs, law, philosophy, and politics, domains that directly link to world beliefs, values, and morality.

We also observe strong performance gains on PRISM for non-US populations, particularly Brazil (68.70% vs. 63.45% for Gemini-2.5-Flash) and India (64.20% vs. 59.80% for Qwen2.5-72B). *PALMs* also achieves the strongest performance in France and Italy. On the US split, Qwen2.5-72B performs slightly better than *PALMs* (62.60% vs. 61.90%). These results suggest that *PALMs* can serve as an effective personalized reward model, generalizing to preference prediction across diverse cultural groups without any task-specific training.

5.3 *Application 2*: Population Simulation

Construct-grounded rationales generalize to behavioral tasks well outside the training distribution. As shown in Fig. 4, *PALMs* achieves the best simulation alignment on 14 out of 17 SIMBENCH datasets, spanning opinion, personality, moral reasoning, and decision-making. Across all datasets, *PALMs* also outperforms Qwen2.5-72B on average (29.62 vs. 27.61).

Gains are largest on construct-aligned tasks (†). Compared with generic rationales, our full method improves by 7.3 points on *OpinionQA*, 18.1 points on *ISSP*, 23.8 points on *OSPsychMACH*, and 16.8 points on *MoralMachine*. These results highlight that construct-grounded rationales help capture latent structure in how people make ethical trade-offs under uncertainty and how deeper personality dimensions, such as Machiavellianism, shape preference judgments.

Beyond construct-aligned tasks, *PALMs* also generalizes to subjective human judgment tasks. On *DICES*, which captures demographic variability in safety assessments of conversational AI responses, *PALMs* achieves a 2.0-point gain over the strongest baseline, suggesting that construct-grounded representations of human diversity also transfer to safety-perception judgments. Similarly, on *TISP*, which measures trust in science and science-related populism, *PALMs* yields a 2.1-point improvement, indicating that value- and belief-grounded rationales generalize to perception-based tasks beyond those directly covered by our construct families. On crowd-knowledge tasks such as *WisdomOfCrowds* and *OSPsychMGKT*, *PALMs* also shows consistent gains of 4.0 and 3.7 points, respectively, indicating that construct rationales help in tasks relying on human expertise.

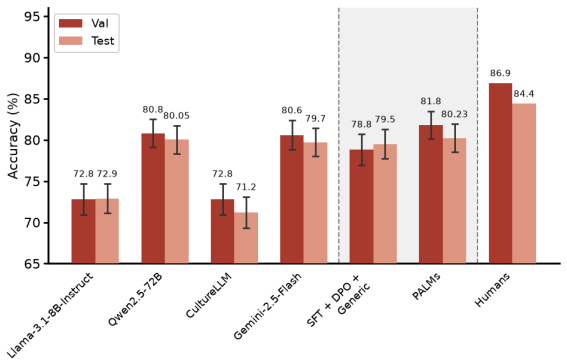

Figure 5: Zero-shot performance (accuracy % with 95% confidence intervals) on the SOCIALIQA validation and test splits. *PALMs* achieves 80.23% test accuracy, slightly outperforming Qwen2.5-72B at 80.05%, and demonstrates strong transfer to social intelligence tasks.

5.4 *Case Study*: Social Intelligence

Construct-grounded rationales improve social intelligence, with particular gains in understanding user intents and motivations. Figure 5 presents zero-shot performance on SOCIALIQA as a case study of whether construct-grounded rationales generalize to social intelligence beyond the preference learning objective. SOCIALIQA requires reasoning about human intent, emotion, and interpersonal expectations making it a natural test of whether construct-grounded rationales transfer beyond preference prediction.

PALMs achieves 80.23% accuracy on the test split, narrowly outperforming Qwen2.5-72B at 80.05% and Gemini-2.5-Flash at 79.72%, despite being an 8B model with no task-specific supervision. On the validation split, *PALMs* achieves 81.80%, compared with 80.80% for Qwen2.5-72B and 80.60% for Gemini-2.5-Flash. *CultureLLM* struggles to perform well on this task (71.18% test accuracy), consistent with patterns observed in the in-domain evaluation, where models trained on

user-level responses struggle to generalize.

We further observe performance gains across various social intelligence question categories (see Tables 9 and 10). *PALMs* shows the largest gains on categories that require reasoning about human intent and motivation. For example, it improves over the strongest baseline included in the category-level analysis by 3.33 points on *motivations*, suggesting that construct-grounded rationales capture meaningful information about why people behave as they do. Although all evaluated models remain below human-level performance (84.40%), these results suggest that construct-grounded rationales transfer to broader social reasoning tasks.

6 Conclusions

In this work, we introduced *Population Aligned Language Models* (*PALMs*), a framework for population-level modeling that uses construct-grounded rationales as latent supervision during preference optimization. Our central finding is that rationalizing through established psychological and cultural constructs provides a richer inductive signal for preference learning than demographic conditioning or survey-based training data alone. Across five culturally diverse populations, *PALMs* consistently improves population alignment, and these gains transfer to personalized reward modeling, population simulation, and social reasoning without any task-specific supervision. We hope this work encourages further exploration of cultural theories in alignment research and opens new directions in multilingual, fine-grained population modeling.

7 Limitations

Country as a Proxy for Population. Following a long line of NLP work that uses country-level annotations as a proxy for cultural population (Li et al., 2024c; Khanuja et al., 2024; Liu et al., 2025b), we adopt country-level abstraction as our unit of analysis. While countries are not monoliths, this granularity is motivated by both conceptual and practical considerations: nations represent long-standing units of shared history, institutional influence, and social norms, and country-level geotagging remains the most common form of demographic partitioning in large-scale datasets. That said, this abstraction inevitably flattens meaningful within-country variation — regional, linguistic, and socioeconomic differences that a single population model cannot capture. Future work should explore finer-grained population definitions, including regional and individual-level modeling.

Choice of Psychological Constructs. The five construct families used in this work — personality traits (OCEAN), cultural dimensions (Hofstede), human values (Schwartz), moral foundations, and primal world beliefs, represent a principled but non-exhaustive selection from a broader landscape of psychological and cultural theory. Other well-established frameworks (Howe, 2012; Brockner and Higgins, 2001) may capture additional dimensions of human preference that our current construct space does not cover. The choice of these five was motivated by their complementary coverage of individual-level psychology and group-level cultural variation, as well as their prevalence in prior NLP work. Future work should explore whether alternative or expanded construct sets yield further improvements in population alignment.

Single Dataset. *Population Aligned Language Models* is trained and evaluated on the Community Alignment dataset, which, while large-scale and culturally diverse, represents one particular operationalization of human preference. Other opinion-based resources, such as social media corpora or alternative survey datasets, may surface different dimensions of population variation. Future work can further explore broader data sources and alternative representations of user identity beyond structured demographic profiles.

Multilingual Modeling. To isolate the effect of construct-grounded reasoning on population alignment without introducing language as a confounding variable, we translated all non-English training instances to English. This means *Population Aligned Language Models* does not model native-language preference expression, which may itself carry culturally meaningful signal. Extending *PALMs* to natively multilingual settings is an important direction for future work.

Model Scale and Architecture. All *Population Aligned Language Models* models are built on `Llama-3.1-8B-Instruct`. While our results demonstrate that multi-construct rationales are effective at this scale, it remains an open question how construct-grounded preference learning interacts with larger models or different base architectures. Future work should examine whether these gains scale and generalize across model families.

References

Shayan Alipour, Indira Sen, Mattia Samory, and Tanu Mitra. 2025. Robustness and confounders in the demographic alignment of llms with human perceptions of offensiveness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22025–22047.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Joel Brockner and E Tory Higgins. 2001. Regulatory focus theory: Implications for the study of emotions at work. *Organizational behavior and human decision processes*, 86(1):35–66.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Vinayak. 2025. Pal: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations*, volume 2025, pages 15755–15799.

Paul M Churchland. 2013. Folk psychology and the explanation of human behavior 1. In *Folk psychology and the philosophy of mind*, pages 247–262. Psychology Press.

Jeremy DW Clifton. 2020. Testing if primal world beliefs reflect experiences—or at least some experiences identified ad hoc. *Frontiers in Psychology*, 11:1145.

Priyanka Dey, Aayush Bothra, Yugal Khanter, Jieyu Zhao, and Emilio Ferrara. 2025. Can LLMs express personality across cultures? introducing CulturalPersonas for evaluating trait alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20241–20262, Suzhou, China. Association for Computational Linguistics.

Priyanka Dey, Daniele Rosa, Wenqing Zheng, Daniel Barcklow, Jieyu Zhao, and Emilio Ferrara. 2026. Gravity: A framework for personalized text generation via profile-grounded synthetic preferences. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7416–7436.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Lewis R. Goldberg. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In *Personality Psychology in Europe*, volume 7, pages 7–28. Tilburg University Press.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *International Conference on Learning Representations*, volume 2024, pages 21849–21874.

Christian W. Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey: Round seven – country-pooled datafile version 5.0.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications.

David Howe. 2012. Attachment theory. *Social work theories and methods*, 75.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2025a. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander Van Der Linden, and Jon Roozenbeek. 2025b. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919.

Oliver John. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. *Published as*.

John A. Johnson. 2014. Johnson's ipip-neo data repository. Accessed: 2024-11-13.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. 2025. Improving language model personas via rationalization with psychological scaffolds. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China. Association for Computational Linguistics*.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Rik Koncel-Kedziorski, Brihi Joshi, and Tim Paek. 2025. PrimeX: A dataset of worldview, opinion, and explanation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24736–24761, Suzhou, China. Association for Computational Linguistics.

Stefan Krsteski, Giuseppe Russo, Serina Chang, Robert West, and Kristina Gligorić. 2025. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. *arXiv preprint arXiv:2510.11408*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024c. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Mengdi Li, Guanqiao Chen, Xufeng Zhao, Haochen Wen, Shu Yang, and Di Wang. 2025. Persrm-r1: Enhance personalized reward modeling with reinforcement learning. *Preprint*, arXiv:2508.14076.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and 1 others. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.

Ziyi Liu, Priyanka Dey, Jen-tse Huang, Zhenyu Zhao, Bowen Jiang, Rahul Gupta, Yang Liu, Yao Du, and Jieyu Zhao. 2025b. Can llms grasp implicit cultural

values? benchmarking llms' cultural intelligence with cq-bench. *arXiv preprint arXiv:2504.01127*.

Qiyao Ma, Dechen Gao, Rui Cai, Boqi Zhao, Hanchu Zhou, Junshan Zhang, and Zhe Zhao. 2026. Personalized rewardbench: Evaluating reward models with human aligned personalization. *arXiv preprint arXiv:2604.07343*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5454–5459.

Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2775–2779.

Nicole Meister, Carlos Guestrin, and Tatsunori B Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49.

Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. 2026. Should you use llms to simulate opinions? quality checks for early-stage deliberation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39070–39079.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, and 1 others. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th annual ACM symposium on user interface software and technology*, pages 1–18.

Pew Research Center. 2023. Spring 2023 global attitudes survey. `https://www.pewresearch.org/global/datasets/`. Survey of 27,285 adults in 23 countries, conducted Feb. 20–May 22, 2023.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8045–8078, Vienna, Austria. Association for Computational Linguistics.

Lilach Sagiv and Shalom H Schwartz. 2022. Personal values across cultures. *Annual review of psychology*, 73:517–546.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25:1–65.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua Yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Conference on Empirical Methods in Natural Language Processing*.

Taylor Sorensen, Jared Moore, Jillian R. Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment. *ArXiv*, abs/2402.05070.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170, Vienna, Austria. Association for Computational Linguistics.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Carlos J Torelli and Andrew M Kaikati. 2009. Values as predictors of judgments and behaviors: The role of abstract and concrete mindsets. *Journal of personality and social psychology*, 96(1):231.

Christin-Melanie Vauclair, Ronald Fischer, Maria Cristina Ferreira, Valeschka Guerra, Ulrich Hößler, Serdar Karabati, Moises Kirk de Carvalho Filho, Juliana B Porto, Melissa Lopez Reyes, Jenni Rytkönen, and 1 others. 2015. What kinds of value motives guide people in their moral attitudes? the role of personal and prescriptive values at the culture level and individual level. *Journal of Cross-Cultural Psychology*, 46(2):211–228.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Lily H Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. 2026. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *The Fourteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Dataset Statistics

A.1 Training Data

We construct preference pairs from country-specific opinion data from the Community Alignment dataset, with one *Population Aligned Language Models* instance trained per population. Table 3 summarizes the size of the training set for each country. Each preference pair consists of a persona, a query, a chosen response, and a rejected response, augmented with a multi-construct rationale (§3.2) for *Population Aligned Language Models* and a generic rationale for the GENERIC baseline. Non-English instances (Hindi, Portuguese, French, and Italian) were translated to English using `Qwen2.5-7B-Instruct` to allow for controlled comparisons across populations without introducing language as a confounding variable; we leave multilingual modeling as future work.

Population	# Preference Pairs
USA	105,999
India	192,696
Brazil	94,380
France	92,348
Italy	97,252

Table 3: Training preference pairs per population.

A.2 Evaluation Datasets

We evaluate *Population Aligned Language Models* along four dimensions of population alignment: personality, cultural norms, values/beliefs, and morality. Table 4 summarizes the number of evaluation items per population for each benchmark.

Personality. We use the IPIP-120 inventory, which contains 120 items spanning the five OCEAN traits (24 items per trait). The same 120 items are used across all populations.

Culture. We use country-specific subsets of CultureBank (Shi et al., 2024); we filter by demographic groups native to each of the five countries. Item counts vary substantially across populations (USA: 1169, Brazil: 78, India: 59, Italy: 459, and France: 329); we therefore report **macro-averaged** accuracy across populations to avoid letting the larger subsets dominate the aggregate score.

Values and beliefs. We aggregate questions from the PEW Global Attitudes Survey across multiple years (2018–2024), filtering to country-specific items. This results in: 86 questions for USA, 51 for Brazil, 74 for France, 51 for India, and 71 for Italy.

Morality. We use questions directly from Wave 7 which are linked with Morality (total: 23). These items were not used to train each CultureLLM model.

Pop.	IPIP	CultureBank	PEW	Morality
USA	120	1,169	86	23
India	120	59	51	23
Brazil	120	78	51	23
France	120	329	74	23
Italy	120	459	71	23

Table 4: Number of evaluation items per population across the four population-alignment benchmarks.

B Training Details

B.1 Compute

All runs are performed on NVIDIA RTX A6000 GPUs (48GB) and NVIDIA A100 (80GB) GPUs. The CULTURELLM baseline uses LoRA (Hu et al., 2022) to fine-tune `Llama-3.1-70B-Instruct` on a single A100 (approximately 1 hour). *Population Aligned Language Models* uses an offline, open-source pipeline: construct-grounded rationales are generated once per population using `Qwen2.5-7B-Instruct`, requiring approximately 16–18 hours per population ($\sim$100K instances). For *Population Aligned Language Models*, we then train full-parameter SFT of `Llama-3.1-8B-Instruct` on 4 RTX A6000 GPUs, taking approximately 12–14 hours per population. DPO is performed on 2$\times$ NVIDIA A100 (80GB) GPUs and takes approximately 12–14 hours per population. We adopt DPO over alternatives such as GRPO-style optimization, as it provides the most direct evaluation of our rationale-augmented preference optimization approach; extending *Population Aligned Language Models* to GRPO-style training is an interesting direction for future work.

B.2 SFT

We perform full-parameter supervised fine-tuning of `Llama-3.1-8B-Instruct` for 2 epochs with the AdamW optimizer, learning rate $2 \times e^{-5}$ with linear warmup (3%), weight decay 0.01, and bf16 precision. We use a per-device batch size of 4 with gradient accumulation of 2 and a maximum sequence length of 1024. Training uses DeepSpeed ZeRO with gradient checkpointing.

B.3 DPO

After SFT, we run Direct Preference Optimization (Rafailov et al., 2023) on the same preference pairs for 2 epochs, with multi-construct rationales included in the prompt context and masked from the loss objective. We use sigmoid DPO with $\beta = 0.1$ and label smoothing 0, the AdamW optimizer with learning rate $2 \times e^{-6}$ on a cosine schedule with 10% warmup, weight decay 0.05, and bf16 precision. We use a per-device batch size of 2 (training) / 4 (evaluation) with gradient accumulation of 4. Training uses DeepSpeed ZeRO with gradient checkpointing.

C Rationale Generation Prompts

We generate multi-construct rationales by prompting an LLM with a shared system message and a construct-specific user message. The system message provides the JSON output format, the persona, the question–answer pair, the ground-truth label, and the general reasoning instructions. The user message then specifies the construct ψ that the rationale should be grounded in. Table 11 lists all six variants used in our experiments: a baseline (no construct) and one for each of the five construct families.

D Example Synthetic Rationales for Training

To illustrate the qualitative difference between generic rationales and multi-construct rationales used during training, we present one preference instance per population (USA, Brazil, India, France, Italy) along with both rationale types for chosen and rejected answers. Each multi-construct rationale decomposes the preference decision across the five construct families introduced in §3.2: personality (OCEAN), cultural dimensions (Hofstede), human values (Schwartz), primal world beliefs, and moral foundations. Tables 12–16 list the full instance and rationales for each population.

E Human Validation of Rationale Quality

To validate the quality of our generated rationales, we conduct a human evaluation study with two annotators who are well-versed in the five psychological and cultural construct families used in this work.

Sample. We randomly sample 50 preference instances, stratified across the five populations (10 per country). For each instance, annotators see the persona, query, candidate answers, and six rationales for both the chosen and rejected answers: one generic rationale and one rationale per construct family (personality, cultural dimensions, human values, primal world beliefs, and moral foundations). This yields a total of $50 \times 6 \times 2 = 600$ rationale ratings per annotator.

Rating rubric. Annotators rate each rationale on a 5-point Likert scale according to how well it invokes its assigned construct (or, for generic rationales, how well it captures relevant aspects of the persona):

- **1**: The construct is not invoked, or is misapplied in a way that contradicts the persona.
- **2**: The construct is mentioned superficially without grounding in the persona's traits or context.
- **3**: The construct is invoked correctly but the reasoning is shallow or partially generic.
- **4**: The construct is invoked correctly and substantively connected to the persona.
- **5**: The construct is accurately and richly invoked, with reasoning clearly grounded in the persona.

Inter-annotator agreement. Table 5 reports Cohen's κ between the two annotators, computed separately for each rationale type. All scores fall in the substantial-agreement range, with an average κ of 0.66 across the six rationale types. The slightly lower agreement for values (0.61) and culture (0.62) reflects the greater subjectivity in deciding when a Schwartz value or Hofstede dimension is "substantively" invoked, while morality shows the highest agreement (0.71), consistent with the more clearly delineated boundaries of the moral foundations.

Rationale type	Cohen's κ
Generic	0.66
Personality (OCEAN)	0.68
Cultural dimensions	0.62
Human values	0.61
Primal world beliefs	0.68
Moral foundations	0.71
Average	**0.66**

Table 5: Inter-annotator agreement (Cohen's κ) on the 5-point quality rubric, computed separately per rationale type.

F Example Model-Generated Rationales at Test Time

To complement the synthetic training-time rationales shown in §D, we present examples of rationales *generated by the trained models at inference time*. For each population, we show one held-out preference instance along with the rationale produced by two trained models: GENERIC-*Population Aligned Language Models* (trained with single-construct generic rationales) and *Population Aligned Language Models* (trained with multi-construct rationales). The same instance is used for both rationale types to facilitate direct comparison. Tables 17–21 list the full instance and rationales for each population. These examples illustrate that the inductive bias from training-time rationale supervision carries over to inference: *Population Aligned Language Models* produces decompositions across the five construct families even though it is given only the persona, and query at test time.

G Ablations

To test whether *PALMs*'s gains stem from the structured construct space itself, rather than simply from additional context, we conduct a leave-one-construct-out ablation on the USA model, removing one construct family's rationales at a time while holding all other training details fixed.

Model	Avg.	Pers.	Cult.	Val.	Mor.
PALMs	**0.726**	**0.656**	**0.772**	**0.672**	**0.802**
w/o Personality	0.714	0.628	0.755	0.661	0.789
w/o Culture	0.708	0.649	0.724	0.655	0.792
w/o Values	0.715	0.650	0.751	0.641	0.793
w/o Morality	0.711	0.651	0.758	0.663	0.773
w/o World Beliefs	0.719	0.653	0.762	0.667	0.793

Table 6: Leave-one-construct-out ablation on the USA PALM model. Removing any single construct family degrades overall alignment, indicating complementary rather than redundant supervision.

Removing any individual construct family consistently degrades performance relative to the full construct space, supporting the claim that the five constructs provide complementary rather than redundant supervision.

H Additional Results

We provide finer-grained breakdowns of the three downstream evaluations summarized in the main paper: personalized reward modeling (PRB), population simulation (SIMBENCH), and social reasoning (SOCIALIQA).

H.1 Personalized Reward Modeling: Domain Breakdown

Table 7 reports PRB accuracy broken down by the three top-level domain categories: *Art*, *Lifestyle*, and *Society*. *Population Aligned Language Models* achieves the best score in every category, with the largest margin on *Society* (+4.45 points over the strongest non-*Population Aligned Language Models* baseline), where preference judgments most heavily depend on cultural and moral reasoning. The SFT + DPO + GENERIC ablation underperforms even the zero-shot `Llama-3.1-8B-Instruct` baseline across all domains, consistent with our broader finding that behavioral fine-tuning without construct-grounded rationales can collapse rather than enrich preference representations.

H.2 Population Simulation: Per-Dataset Breakdown

Table 8 reports per-dataset SIMBENCH scores, macro-averaged across countries. Scores are signed (higher is better; 0 corresponds to random; negative values indicate worse-than-random simulation). *Population Aligned Language Models* achieves the best score on the majority of datasets and is competitive on the remainder. The largest gains occur on construct-aligned datasets (†): on OPINIONQA, GLOBOPINIONQA, ISSP, OSPSYCHBIG5, and OSPSYCHMACH, *Population Aligned Language Models* substantially outperforms all baselines—most notably on OSPSYCHMACH, where every other method scores below random while *Population Aligned Language Models* reaches 14.6. This pattern supports our central claim: when downstream tasks probe the same psychological and cultural constructs that shape *Population Aligned Language Models*'s training-time rationales, the construct-grounded representation transfers strongly. On non-construct-aligned datasets such as CHAOSNLI and CONSPIRACYCORR, *Population Aligned Language Models* performs comparably to GEMINI-2.5-FLASH despite using a much smaller base model.

H.3 Social Reasoning: SOCIALIQA Breakdown

Tables 9 and 10 report fine-grained zero-shot accuracy on the SOCIALIQA validation and test splits

Model	Art	Lifestyle	Society	Overall
Skywork-Reward-V2	$63.10_{\pm 3.25}$	$65.20_{\pm 2.75}$	$64.90_{\pm 2.60}$	$64.40_{\pm 1.68}$
Llama-3.1-8B-Instruct	$60.63_{\pm 3.45}$	$65.02_{\pm 2.97}$	$61.24_{\pm 2.91}$	$62.29_{\pm 1.79}$
Qwen2.5-72B	$62.55_{\pm 3.30}$	$64.25_{\pm 2.82}$	$63.40_{\pm 2.71}$	$63.40_{\pm 1.72}$
CultureLLM	$60.15_{\pm 3.46}$	$65.86_{\pm 2.95}$	$58.97_{\pm 2.94}$	$61.64_{\pm 1.79}$
Gemini-2.5-Flash	$63.15_{\pm 3.19}$	$65.33_{\pm 2.68}$	$64.90_{\pm 2.53}$	$64.46_{\pm 1.60}$
PAL	$62.90_{\pm 3.22}$	$64.10_{\pm 2.76}$	$64.70_{\pm 2.61}$	$63.90_{\pm 1.68}$
SFT + DPO + Generic	$52.10_{\pm 3.20}$	$54.40_{\pm 2.85}$	$53.25_{\pm 2.90}$	$53.25_{\pm 1.75}$
***Population Aligned Language Models* (Ours)**	$\mathbf{65.20}_{\pm 3.10}$	$\mathbf{66.12}_{\pm 2.60}$	$\mathbf{69.35}_{\pm 2.45}$	$\mathbf{66.89}_{\pm 1.65}$

Table 7: Personalized reward modeling accuracy (%) on the PRB benchmark, broken down by domain category. Values are mean $\pm$ bootstrapped 95% confidence interval margin.

Dataset	Llama-3.1-8B	Qwen2.5-72B	CultureLLM	Gemini-2.5-Flash	SFT+DPO+Generic	*PALMs* (Ours)
ChaosNLI	$-22.16_{\pm 7.23}$	$\mathbf{44.80}_{\pm 4.35}$	$-9.65_{\pm 6.35}$	$40.64_{\pm 4.83}$	$38.20_{\pm 4.50}$	$43.60_{\pm 4.10}$
Choices13k	$-130.76_{\pm 10.83}$	$-19.20_{\pm 6.50}$	$-101.70_{\pm 11.38}$	$-25.27_{\pm 7.18}$	$-24.10_{\pm 6.80}$	$\mathbf{-17.70}_{\pm 6.20}$
ConspiracyCorr	$-212.74_{\pm 71.83}$	$17.30_{\pm 27.00}$	$-177.86_{\pm 52.30}$	$\mathbf{22.08}_{\pm 31.23}$	$6.40_{\pm 28.50}$	$16.50_{\pm 25.10}$
DICES	$68.66_{\pm 4.26}$	$73.30_{\pm 3.65}$	$56.42_{\pm 5.03}$	$66.12_{\pm 4.06}$	$69.50_{\pm 3.80}$	$\mathbf{75.30}_{\pm 3.50}$
ESS	$-44.94_{\pm 32.91}$	$36.10_{\pm 12.90}$	$-19.55_{\pm 28.29}$	$29.56_{\pm 14.74}$	$31.00_{\pm 13.50}$	$\mathbf{38.50}_{\pm 12.10}$
GlobOpinionQA†	$-20.34_{\pm 16.00}$	$49.50_{\pm 8.75}$	$-12.59_{\pm 15.45}$	$32.39_{\pm 10.02}$	$35.20_{\pm 9.50}$	$\mathbf{52.50}_{\pm 8.20}$
ISSP†	$-11.42_{\pm 16.09}$	$45.70_{\pm 10.20}$	$2.94_{\pm 15.02}$	$27.59_{\pm 12.49}$	$30.10_{\pm 11.20}$	$\mathbf{48.20}_{\pm 9.80}$
Jester	$-189.81_{\pm 1.37}$	$-36.50_{\pm 5.35}$	$-132.94_{\pm 11.86}$	$-42.61_{\pm 6.35}$	$-41.50_{\pm 5.80}$	$\mathbf{-35.10}_{\pm 5.10}$
LatinoBarometer	$-52.60_{\pm 26.82}$	$-25.50_{\pm 20.30}$	$-35.03_{\pm 25.15}$	$-32.83_{\pm 26.56}$	$-31.50_{\pm 22.40}$	$\mathbf{-24.20}_{\pm 19.80}$
MoralMachine†	$-31.75_{\pm 13.27}$	$-12.30_{\pm 10.00}$	$-29.34_{\pm 12.17}$	$-85.54_{\pm 15.56}$	$-25.10_{\pm 11.00}$	$\mathbf{-8.30}_{\pm 9.50}$
NumberGame	$-17.87_{\pm 8.38}$	$\mathbf{23.70}_{\pm 5.15}$	$-5.99_{\pm 6.16}$	$16.43_{\pm 6.14}$	$19.80_{\pm 5.50}$	$23.20_{\pm 4.90}$
OSPsychBig5†	$6.28_{\pm 11.18}$	$40.70_{\pm 7.65}$	$5.93_{\pm 9.68}$	$23.28_{\pm 9.49}$	$26.40_{\pm 8.50}$	$\mathbf{42.90}_{\pm 7.20}$
OSPsychMACH†	$-11.15_{\pm 24.78}$	$15.10_{\pm 19.60}$	$-10.08_{\pm 38.55}$	$-32.16_{\pm 27.35}$	$-5.20_{\pm 22.10}$	$\mathbf{18.60}_{\pm 18.50}$
TISP	$-33.61_{\pm 22.88}$	$24.40_{\pm 13.30}$	$-11.96_{\pm 20.23}$	$18.34_{\pm 16.09}$	$19.80_{\pm 14.50}$	$\mathbf{26.50}_{\pm 12.80}$
OpinionQA†	$30.88_{\pm 3.68}$	$65.50_{\pm 1.98}$	$34.31_{\pm 3.20}$	$59.35_{\pm 2.28}$	$61.20_{\pm 2.10}$	$\mathbf{68.50}_{\pm 1.85}$
WisdomOfCrowds	$4.87_{\pm 11.88}$	$64.10_{\pm 5.85}$	$20.74_{\pm 9.57}$	$56.35_{\pm 6.72}$	$58.20_{\pm 6.20}$	$\mathbf{68.10}_{\pm 5.50}$
OSPsychMGKT	$13.50_{\pm 17.16}$	$62.70_{\pm 8.70}$	$40.00_{\pm 12.98}$	$51.32_{\pm 10.21}$	$53.50_{\pm 9.50}$	$\mathbf{66.40}_{\pm 8.20}$
Average	-38.53	27.61	-22.73	13.24	18.94	$\mathbf{29.62}$

Table 8: Population simulation on SIMBENCH. Higher is better; 0 indicates random performance and negative values indicate performance below random. Scores are macro-averaged across countries and include 95% confidence interval margins. † denotes construct-aligned datasets.

across its five reasoning categories: *Wants*, *Reactions*, *Descriptions*, *Motivations*, and *Needs*. *Population Aligned Language Models* achieves the best score on every category and overall, narrowing the gap to the human baseline (86.9 on validation, 84.4 on test) more than any other approach we evaluate. The strongest gains are on *Motivations*—the category most directly tied to underlying values, beliefs, and moral intuitions—where *Population Aligned Language Models* improves over `Llama-3.1-8B-Instruct` by 5.6 points on validation and 6.4 points on test. This suggests that construct-grounded rationales transfer beyond preference prediction to general social-reasoning tasks, even without task-specific supervision.

Model	Wants	React.	Descr.	Motiv.	Needs	Overall
Llama-3.1-8B-Instruct	$69.13_{\pm 3.84}$	$75.18_{\pm 4.18}$	$73.17_{\pm 5.17}$	$78.37_{\pm 5.13}$	$71.17_{\pm 5.92}$	$72.82_{\pm 1.97}$
Qwen2.5-72B	$80.50_{\pm 3.28}$	$82.10_{\pm 3.62}$	$79.80_{\pm 4.58}$	$82.60_{\pm 4.70}$	$79.00_{\pm 5.18}$	$80.80_{\pm 1.70}$
CultureLLM	$69.13_{\pm 3.84}$	$77.15_{\pm 4.07}$	$72.47_{\pm 5.14}$	$78.78_{\pm 5.10}$	$67.57_{\pm 6.11}$	$72.77_{\pm 1.97}$
SFT + DPO + Generic	$77.20_{\pm 3.50}$	$79.10_{\pm 3.85}$	$78.40_{\pm 4.80}$	$80.20_{\pm 4.90}$	$79.10_{\pm 5.30}$	$78.80_{\pm 1.80}$
Gemini-2.5-Flash	$79.42_{\pm 3.36}$	$82.31_{\pm 3.70}$	$79.09_{\pm 4.69}$	$82.04_{\pm 4.79}$	$80.18_{\pm 5.22}$	$80.60_{\pm 1.75}$
***Population Aligned Language Models* (Ours)**	$\mathbf{81.35}_{\pm 3.20}$	$\mathbf{82.40}_{\pm 3.55}$	$\mathbf{80.20}_{\pm 4.50}$	$\mathbf{83.95}_{\pm 4.60}$	$\mathbf{81.10}_{\pm 5.10}$	$\mathbf{81.80}_{\pm 1.65}$
Human Baseline	86.40	87.10	86.50	87.90	86.60	86.90

Table 9: Fine-grained zero-shot accuracy (%) on the SOCIALIQA **validation split** across the five construct categories. Values are mean ± bootstrapped 95% confidence interval margin.

Model	Wants	React.	Descr.	Motiv.	Needs	Overall
Llama-3.1-8B-Instruct	$72.38_{\pm 3.62}$	$72.95_{\pm 3.89}$	$74.41_{\pm 4.94}$	$74.83_{\pm 4.91}$	$72.43_{\pm 5.28}$	$72.89_{\pm 1.85}$
Qwen2.5-72B	$80.90_{\pm 3.18}$	$79.50_{\pm 3.48}$	$80.80_{\pm 4.38}$	$77.50_{\pm 4.62}$	$81.55_{\pm 4.48}$	$80.05_{\pm 1.70}$
CultureLLM	$72.04_{\pm 3.63}$	$72.34_{\pm 3.91}$	$72.05_{\pm 5.08}$	$73.15_{\pm 5.01}$	$66.91_{\pm 5.56}$	$71.18_{\pm 1.88}$
SFT + DPO + Generic	$78.80_{\pm 3.35}$	$79.10_{\pm 3.60}$	$79.50_{\pm 4.60}$	$80.20_{\pm 4.80}$	$79.90_{\pm 4.75}$	$79.50_{\pm 1.70}$
Gemini-2.5-Flash	$79.93_{\pm 3.25}$	$79.56_{\pm 3.53}$	$80.81_{\pm 4.47}$	$77.85_{\pm 4.70}$	$81.99_{\pm 4.56}$	$79.72_{\pm 1.67}$
***Population Aligned Language Models* (Ours)**	$\mathbf{81.60}_{\pm 3.10}$	$\mathbf{79.70}_{\pm 3.40}$	$\mathbf{81.15}_{\pm 4.30}$	$\mathbf{81.18}_{\pm 4.50}$	$\mathbf{82.10}_{\pm 4.40}$	$\mathbf{80.23}_{\pm 1.55}$
Human Baseline	84.20	84.50	84.10	85.00	84.20	84.40

Table 10: Fine-grained zero-shot accuracy (%) on the SOCIALIQA **test split** across the five construct categories. Values are mean ± bootstrapped 95% confidence interval margin.

System Message:

Please generate your response in the following valid JSON format:

```
{
  "reasoning_chain": ["Step 1: ...", "Step 2: ...", "..."],
  "conclusion": "chosen" or "not chosen"
}
```

You will be provided with: a persona, a question and candidate answer, and the correct label. Reason step-by-step using the construct below to predict and justify whether this answer would be chosen by this persona. Each step should: (1) make a concrete evaluative claim, (2) build on the previous step, (3) explicitly connect persona traits to the answer's content via the construct. The final step must commit to a prediction. The conclusion must match the label provided.

Construct: `$construct` **Persona:** `$persona` **Question:** `$question` **Answer:** `$answer` **Label:** `$label`

User Message:

BASELINE (NO CONSTRUCT)

Focus on the user's background, preferences, and opinions to explain the reasoning chain. Generate 4–6 steps, each highlighting a different aspect of the persona. Where possible, show how multiple aspects interact to support the correct answer.

PERSONALITY (OCEAN)

Focus on the user's personality profile as defined by the OCEAN model:

- *Openness to Experience*: curiosity, imagination, and openness to new ideas, experiences, and perspectives.
- *Conscientiousness*: organization, dependability, discipline, and goal-orientation.
- *Extraversion*: outgoingness, energy, and drawing energy from social interactions.
- *Agreeableness*: cooperativeness, empathy, trust, and consideration for others.
- *Neuroticism*: emotional instability, anxiety, moodiness, or negative affect.

Generate 4–6 steps, each focusing on a different trait or combination of traits. If traits interact in a way that informs the persona's reasoning, include that interaction explicitly.

CULTURAL DIMENSIONS (HOFSTEDE)

Focus on the user's cultural profile as defined by Hofstede's Cultural Dimensions Theory:

- *Individualism vs. Collectivism*: prioritizing personal goals and autonomy vs. group loyalty and collective well-being.
- *Power Distance*: acceptance of unequal distribution of power in society and institutions.
- *Uncertainty Avoidance*: discomfort with ambiguity and preference for rules, structure, and predictability.
- *Masculinity vs. Femininity*: valuing competition and material success vs. cooperation, care, and quality of life.
- *Long-Term vs. Short-Term Orientation*: focus on future rewards and perseverance vs. tradition and immediate results.
- *Indulgence vs. Restraint*: enjoying life and acting on impulses vs. suppressing gratification through social norms.

Generate 4–6 steps, each highlighting a different cultural dimension or their interaction.

HUMAN VALUES (SCHWARTZ)

Focus on the user's values as defined by Schwartz's Theory of Basic Human Values:

- *Self-Direction*: independent thought, creativity, and freedom to choose one's own goals.
- *Stimulation*: novelty, excitement, and a varied, challenging life.
- *Hedonism*: pleasure, sensual gratification, and personal enjoyment.
- *Achievement*: personal success and demonstrating competence according to social standards.
- *Power*: social status, prestige, and control over people and resources.
- *Security*: safety, harmony, and stability of society, relationships, and self.
- *Conformity*: restraint of actions and impulses that might upset others or violate social norms.
- *Tradition*: respect for and commitment to the customs and ideas of one's culture or religion.
- *Benevolence*: preservation and enhancement of the welfare of people one is close to.
- *Universalism*: understanding, appreciation, tolerance, and protection for the welfare of all people and nature.

Generate 4–6 steps showing a different value or interplay between values that affects the persona's reasoning.

MORAL FOUNDATIONS

continued on next page

Table 11 continued from previous page.

Focus on the user's moral profile as defined by Moral Foundations Theory (MFT):

- *Care/Harm*: sensitivity to suffering and motivation to protect the vulnerable; valuing kindness and compassion.
- *Fairness/Cheating*: concern for justice, rights, and equal treatment; valuing reciprocity and honesty.
- *Loyalty/Betrayal*: valuing solidarity, duty, and self-sacrifice for one's group.
- *Authority/Subversion*: valuing respect for legitimate authority, tradition, and social hierarchy.
- *Sanctity/Degradation*: sense that some things are sacred; valuing purity, cleanliness, and elevation over baseness.

Generate 4–6 steps, highlighting different moral foundations and how they work together to justify the correct answer.

PRIMAL WORLD BELIEFS

Focus on the user's Primal World Beliefs—fundamental beliefs about the overall character of the world:

- *Safe vs. Dangerous*: world seen as benign vs. threatening and hazardous.
- *Enticing vs. Dull*: world full of opportunity and wonder vs. boring and unstimulating.
- *Alive vs. Mechanistic*: world has intention and meaning vs. random, mechanical, and indifferent.
- *Just vs. Unjust*: world is fair and rewards good behavior vs. arbitrary and indifferent to merit.
- *Abundant vs. Barren*: world is full of resources and goodness vs. scarce and withholding.
- *Interconnected vs. Fragmented*: world is a coherent connected whole vs. a collection of unrelated parts.

Generate 4–6 steps covering multiple primal beliefs and how they interact to support the correct answer.

Table 11: **Prompts used to generate multi-construct rationales.** A shared system message specifies the JSON output format and reasoning instructions, with placeholders for the persona, question, candidate answer, and ground-truth label. The user message injects construct-specific instructions ψ, one per construct family.

Table 12: **Example preference instance and training-time rationales for USA.**

Persona. You are a Hispanic or Latino, 18–34 male from the United States with education level: post-secondary graduate and political leaning: very left-leaning.

Question. How do I maintain a consistent routine if I have ADHD?

Chosen Answer. Flexibility is essential when dealing with ADHD. Rather than adhering to a rigid schedule, allow for spontaneity and adjustments as needed. Prioritize tasks based on importance and deadlines, focusing on completing the high-priority tasks first. Using the Pomodoro Technique, which involves working in focused 25-minute increments followed by a five-minute break, can help maintain productivity and avoid burnout. Embracing flexibility and being kind to oneself when deviations from the routine occur is vital for long-term success and reducing stress.

Rejected Answer. Breaking down tasks into smaller, manageable chunks is key. Create a schedule with specific times allocated for each activity, including breaks. Use visual aids like calendars or planners to stay organized and on track. Implementing a "brain dump" session daily can help clear mental clutter, allowing focus on the task at hand. Consistency is crucial, so try to perform tasks at the same time every day to create habits.

GENERIC RATIONALE — CHOSEN

Step 1: The answer emphasizes flexibility and adaptability, which is likely to resonate with a young adult navigating ADHD while balancing multiple responsibilities. Someone in the 18–34 age group may appreciate approaches that feel realistic and less rigid. *Step 2:* The inclusion of practical productivity strategies, such as prioritizing tasks and using the Pomodoro Technique, aligns well with the persona's post-secondary educational background. Individuals with higher education experience are often familiar with structured time-management methods and may find them credible and actionable. *Step 3:* The response balances structure with self-compassion, encouraging the individual to avoid self-criticism when routines are disrupted. This supportive tone may appeal to someone who values mental health awareness and emotional well-being. *Step 4:* The advice avoids overly strict or prescriptive routines and instead promotes personalization and spontaneity. This can feel more relatable and sustainable for a younger individual managing ADHD symptoms in everyday life. *Step 5:* Overall, the answer combines practical guidance with empathy, making it likely to resonate with a young, educated individual seeking manageable and emotionally supportive strategies for maintaining consistency.

GENERIC RATIONALE — REJECTED

Step 1: The answer provides practical organizational strategies such as scheduling tasks, using planners, and maintaining consistency. While useful, the overall approach may feel somewhat rigid for a young adult managing ADHD. *Step 2:* The emphasis on strict routines and performing tasks at the same time every day may not fully align with the needs of someone who benefits from flexibility and adaptability in their daily structure. *Step 3:* Although the response includes actionable suggestions like breaking tasks into smaller chunks and using visual aids, the tone focuses primarily on productivity and organization rather than emotional support or self-compassion. *Step 4:* The answer presents fairly standard self-management advice without much personalization or acknowledgment of the emotional challenges associated with ADHD. A younger individual may prefer guidance that feels more understanding and tailored to real-life experiences. *Step 5:* Overall, while the response is practical and structured, it may feel too conventional and prescriptive compared to approaches that balance productivity with flexibility and mental well-being.

MULTI-CONSTRUCT RATIONALE — CHOSEN

Personality. *Step 1:* The persona, being a young male with a post-secondary education level, likely values practical and implementable strategies. The answer provides actionable advice on maintaining a routine with ADHD, which aligns well with the persona's likely preference for structured solutions. *Step 2:* The emphasis on flexibility in the answer connects to the persona's possible high openness to experience, as it suggests adapting routines rather than following rigid structures. This aligns with someone who may appreciate creative and adaptable approaches to problem-solving. *Step 3:* The use of the Pomodoro Technique reflects a conscientious aspect, promoting organization and goal-oriented behavior. Given the persona's educational background, he might resonate with structured productivity techniques, reinforcing the relevance of the answer. *Step 4:* The recommendation to be kind to oneself when deviations occur demonstrates a high level of agreeableness. This could appeal to the persona, who may value empathy and self-care principles, particularly if he leans towards progressivism in his political views. *Step 5:* The answer addresses the nuances of dealing with ADHD, acknowledging emotional aspects as well. People with ADHD often experience varying emotional states, so the focus on reducing stress and prioritizing mental well-being could resonate strongly with the persona.

Cultural Dimensions. *Step 1:* The answer emphasizes the importance of flexibility when managing ADHD, which aligns well with the persona's inclination towards adaptability in routines, reflecting a practical understanding of individual needs. *Step 2:* By suggesting the Pomodoro Technique, the answer provides a structured method that acknowledges the persona's educational background. A post-secondary graduate may appreciate effective strategies that help with productivity, suggesting that the persona values practical, evidence-based solutions. *Step 3:* The advice to prioritize tasks shows an understanding of the need for effective time management, which resonates with a post-secondary graduate's likely experience in balancing multiple responsibilities. This reflects their potential long-term orientation. *Step 4:* The focus on kindness towards oneself when routines deviate reflects a more feminine approach to managing life's challenges, prioritizing emotional well-being over strict adherence to tasks. Given the persona's left-leaning views, this compassionate perspective is likely to resonate strongly. *Step 5:* The suggestion to embrace spontaneity speaks to the persona's probable inclination towards creativity and personal expression, common among individuals who lean left politically, supporting the idea of personal autonomy while managing ADHD.

continued on next page

Table 12 continued.

Human Values. *Step 1:* The answer emphasizes flexibility, which aligns with the Self-Direction value of the persona. As someone who values independent thought and the freedom to choose their own goals, the persona would likely appreciate the idea of not sticking to a rigid routine but instead adapting to their own needs. *Step 2:* The suggestion to prioritize tasks and use the Pomodoro Technique aligns with the Achievement value. Since the persona is a post-secondary graduate, they likely understand the importance of competence and effectiveness in managing tasks. *Step 3:* The emphasis on being kind to oneself when deviations occur supports the Hedonism value. It recognizes the importance of personal enjoyment and stress reduction, suggesting that the persona may appreciate methods that prioritize mental well-being, particularly given their left-leaning political views which may advocate for compassionate approaches. *Step 4:* The acknowledgment of spontaneity in managing ADHD resonates with the Stimulation value. The persona, valuing novelty and a varied life, might find the idea of adjusting their routine based on current circumstances refreshing and engaging. *Step 5:* Lastly, the holistic perspective of maintaining a balanced life with ADHD reflects the Universalism value. The idea promotes understanding and acceptance of personal challenges, aligning with the persona's potential advocacy for empathy and support.

Primal World Beliefs. *Step 1:* The answer emphasizes flexibility, which aligns well with the persona's likely view of life as complex and fluid due to their experience with ADHD. This suggests that the persona values adaptability in their daily routine. *Step 2:* By advocating for spontaneity and adjustments, the answer connects with the persona's understanding that life is not always predictable. This reflects a belief in a world that is alive and dynamic, resonating with their experiences. *Step 3:* The mention of prioritizing tasks by importance can be tied to the persona's notion of a just and fair approach to productivity. It suggests a structure that rewards good management of time and resources, appealing to their desire for a just world where effort is recognized. *Step 4:* Using the Pomodoro Technique introduces a systematic yet flexible method for managing tasks, reinforcing the persona's potential belief in a world that is abundant with tools and methods to achieve goals. This structure showcases a rich landscape of strategies to enhance productivity. *Step 5:* Embracing kindness toward oneself when routines deviate speaks to a larger narrative of self-acceptance and resilience. This aligns with the persona's perspective that life's challenges should be met with compassion, reinforcing their worldview that life is interconnected and that personal well-being is crucial in the broader context of social justice.

Moral Foundations. *Step 1:* The answer emphasizes flexibility, which aligns well with the Care/Harm moral foundation by promoting a compassionate approach to managing ADHD. This suggests sensitivity towards the challenges of individuals with ADHD, highlighting the importance of self-kindness and understanding. *Step 2:* The mention of prioritizing tasks according to importance and deadlines reflects a sense of Fairness/Cheating, as it suggests a methodical approach to ensuring that one's efforts are aligned with obtaining just rewards for hard work. This aligns with the persona's left-leaning political views, where values of equity are prioritized. *Step 3:* Incorporating the Pomodoro Technique promotes the idea of structured productivity, which is important in maintaining focus amidst distractions. This reflects the Authority/Subversion foundation, as it implies respect for proven methods, resonating with the persona's educated background. *Step 4:* The emphasis on being kind to oneself and allowing for deviations reinforces the Care/Harm foundation through compassionate self-management, indicating a nurturing approach to personal growth. *Step 5:* The overall tone of the response is supportive and empowering, which is likely to appeal to a young, educated individual from a marginalized group, as they may value guidance that is both pragmatic and empathetic.

MULTI-CONSTRUCT RATIONALE — REJECTED

Personality. *Step 1:* The answer provides practical advice on maintaining a routine for those with ADHD, which could appeal to a user who values organization and discipline, aligning with conscientiousness traits. However, the user's inclination to seek creative or alternative strategies due to their openness to experience may find this method too rigid. *Step 2:* The suggestion to break tasks into manageable chunks demonstrates a structured approach. However, the persona, being from a very left-leaning political background, may prefer more innovative and less conventional methods rather than strictly organizational techniques, making this answer feel too conventional. *Step 3:* The answer emphasizes creating a schedule and using visual aids, which caters to organization. Yet, this strict adherence to scheduling may not resonate well with the persona if they value flexibility and creativity in their routines. *Step 4:* The use of "brain dump" sessions indicates a recognition of the mental clutter ADHD can cause, which may appeal to the persona's empathetic side. However, the overall tone of the answer lacks a collaborative or supportive approach, which may clash with the persona's high agreeableness. *Step 5:* The emphasis on consistency may resonate with the user's conscientiousness and desire for stability. However, given their likely experience navigating a world that often doesn't accommodate ADHD, the persona might appreciate more personalized or nuanced strategies.

Cultural Dimensions. *Step 1:* The answer promotes structured approaches like creating schedules and using visual aids. However, the persona's cultural background may value collective rather than individual achievement, making a rigidly individualistic approach less appealing. *Step 2:* Considering the persona's left-leaning political views, there may be a preference for solutions that emphasize community and collaborative support, rather than solitary strategies. This answer lacks a communal focus, which might weaken its appeal. *Step 3:* The suggestion to implement a "brain dump" session implies a certain level of personal introspection and self-reliance that may conflict with a more collectivist mindset, which often prioritizes social and emotional support systems over individual coping mechanisms. *Step 4:* The emphasis on routine and predictability reflects high Uncertainty Avoidance, which might appeal to someone with ADHD. However, the persona may also appreciate flexibility in routines that adapt to social interactions, rather than strictly adhering to a predetermined schedule. *Step 5:* The cultural dimension of Indulgence vs. Restraint is relevant here; the focus on managing tasks may come off as overly restrictive. The persona might prefer a more indulgent approach to life, where enjoyment and adaptability take precedence over rigorous self-management techniques.

continued on next page

Table 12 continued.

Human Values. *Step 1:* The persona is a post-secondary graduate, likely valuing Achievement and Self-Direction, which means they appreciate strategies that demonstrate competence and offer independence in managing ADHD. However, the answer mainly offers organizational tips rather than empowering the persona to develop their unique strategies. *Step 2:* While the response does provide practical solutions like breaking down tasks and using visual aids, it lacks a creative or innovative approach that might resonate with the persona's preference for Stimulation. The answers are quite standard and may not engage their desire for a varied and challenging approach. *Step 3:* The recommendation to create a schedule is useful, but it could be perceived as conforming to a rigid structure, which may conflict with the persona's potential value for flexibility and independent thought. *Step 4:* The suggestion of a "brain dump" could be seen as a step towards personal expression and self-exploration, aligning with Self-Direction; however, the answer overall does not reflect wider considerations for emotional and psychological well-being that may cater to their left-leaning political views, which emphasize social and community support for mental health. *Step 5:* Ultimately, the answer lacks depth and is too prescriptive, offering a one-size-fits-all solution rather than encouraging a more relatable, individualized approach to managing ADHD.

Primal World Beliefs. *Step 1:* The answer focuses on practical strategies for managing ADHD, which seems reasonable. However, it lacks a deeper emotional engagement that could resonate with a young Hispanic or Latino male seeking both empathy and understanding in his struggles with ADHD. *Step 2:* The persona may find the world as a challenging place due to personal or systemic factors, which makes it less enticing to approach tasks merely through structured methods. This reliance on mechanical approaches might not align well with the persona's inclination for a more interconnected and meaningful approach to personal struggles. *Step 3:* The answer emphasizes consistency and organization, which appeal to a sense of safety and control. Yet, for someone who might see the world as more dynamic and full of rich, complex interactions, this 'one-size-fits-all' approach feels somewhat superficial and dismissive of personal experiences tied to cultural identity. *Step 4:* The answer does not address the emotional aspects of ADHD, such as feelings of frustration or isolation. A young man who is Hispanic or Latino, and likely aware of social stigma around mental health, may prefer guidance that acknowledges these feelings and offers support in a more compassionate manner. *Step 5:* Overall, the answer seems to be overly focused on functional strategies without adequately considering the persona's psychological and emotional framework. It misses the opportunity to connect on the level of living an abundant and fulfilling life despite ADHD.

Moral Foundations. *Step 1:* The answer provides practical strategies for managing ADHD, which could be seen as beneficial, but it lacks an emphasis on empathy or support for individuals dealing with this condition, missing the Care/Harm foundation that values kindness. *Step 2:* The advice given is somewhat prescriptive and focused on individual responsibility without acknowledging the systemic issues that may affect individuals with ADHD, such as access to mental health resources, which conflicts with the Fairness/Cheating foundation. *Step 3:* The tone of the answer is clinical and focused on self-management strategies; it does not foster a sense of community or loyalty among individuals experiencing ADHD, lacking the Loyalty/Betrayal foundation. *Step 4:* Furthermore, while the answer suggests creating routines and schedules, it does not address the unique challenges faced by individuals from marginalized backgrounds, where systemic barriers may hinder the ability to implement these strategies effectively. This neglect reveals a lack of sensitivity to context. *Step 5:* The moral framing leans toward individual self-management, which is valid but less aligned with the relational, mentor-respecting moral orientation the persona likely brings.

Table 13: **Example preference instance and training-time rationales for Brazil.**

Persona. You are a 35–45-year-old male from Brazil, post-secondary graduate, somewhat left-leaning, White.

Question (translated to English). Which markets should be visited in Venice?

Chosen Answer (translated). Venice is famous for its traditional markets, where you can find a variety of local products, including fruits, vegetables, fresh fish, and handmade crafts. The Rialto Market is one of the most well-known and is a must-visit destination for any traveler. In addition, the San Polo Market also offers a unique experience, with its vibrant colors and enticing aromas. Visiting these markets not only provides an opportunity to experience local cuisine but also to immerse yourself in the culture and history of the city.

Rejected Answer (translated). In Venice, the most well-known markets include the Rialto Market, famous for fruits, vegetables, and fresh fish, and the San Maurizio Market, which offers a variety of local and handmade products. These markets are great places to experience Venetian cuisine and culture.

GENERIC RATIONALE — CHOSEN

Step 1: The answer provides a rich, descriptive overview of Venice's traditional markets, which is likely to appeal to a mid-career graduate looking for cultural travel guidance that goes beyond surface information. *Step 2:* The narrative style brings the markets to life through sensory detail and atmosphere, fitting a persona who likely values immersive experiences over checklist tourism. *Step 3:* The emphasis on local cuisine, culture, and history adds the kind of depth that resonates with Brazilian cultural appreciation for food, tradition, and meaningful human encounters. *Step 4:* The warm, respectful tone fits the persona's somewhat left-leaning orientation by celebrating local heritage and the human dimension of marketplaces without ideological framing. *Step 5:* Overall, the answer balances practical recommendations with cultural depth, making it a strong fit for a mid-career Brazilian traveler seeking a meaningful Venetian market experience.

GENERIC RATIONALE — REJECTED

Step 1: The answer provides accurate information about two Venetian markets, but its briefer, more clinical structure may feel less engaging for a mid-career graduate seeking richer cultural guidance. *Step 2:* The absence of vivid sensory description underweights the atmosphere and human warmth that often resonate with Brazilian cultural sensibilities and meaningful travel experiences. *Step 3:* The compact treatment provides basic information but lacks the immersive depth that helps a culturally engaged adult connect with a destination. *Step 4:* The neutral tone fits the persona's somewhat left-leaning orientation, but the lighter framing offers less of the thoughtful cultural depth he likely values. *Step 5:* Overall, while the answer is clear and informative, its thinner detail and limited atmospheric framing make it a slightly weaker fit for a mid-career Brazilian traveler seeking a richly evocative Venetian market experience.

MULTI-CONSTRUCT RATIONALE — CHOSEN

Personality. *Step 1:* The persona, as a mid-career Brazilian male graduate, likely exhibits high openness to experience, which aligns with the answer's evocative descriptions of vibrant colors, enticing aromas, and the sensory richness of Venetian markets. The descriptive framing invites him to imagine himself fully immersed in the experience. *Step 2:* The answer's narrative flow reflects a balance between conscientiousness and creativity, providing organized cultural guidance without feeling overly transactional or list-driven. As an educated professional, the persona likely appreciates this thoughtful, well-shaped storytelling approach. *Step 3:* The framing of markets as places to "immerse yourself in the culture and history" taps into a balanced extraversion profile, encouraging social and sensory engagement that resonates with a mid-career traveler likely seeking meaningful, atmosphere-rich experiences. *Step 4:* The warm and inviting tone reflects high agreeableness, treating the markets as living spaces that welcome visitors rather than as commodified attractions. The persona, as a somewhat left-leaning adult, may appreciate this respectful framing of local culture. *Step 5:* The answer reduces decision pressure by foregrounding atmosphere and experience over transactional details, suiting a persona with low to moderate neuroticism who likely prefers travel guidance that feels confident and unhurried rather than overwhelming.

Cultural Dimensions. *Step 1:* The descriptive emphasis on local products, fresh fish, fruits, and handmade crafts reflects an appreciation for cultural authenticity and tradition, which aligns with Brazilian cultural sensibilities that value sensory, communal, and craft-rooted experiences. *Step 2:* The framing of markets as immersive cultural spaces engages an indulgence-oriented orientation, celebrating sensory pleasure, food, and the warmth of marketplace life—values that resonate strongly with both Brazilian culture and the persona's likely appreciation for richly lived experiences. *Step 3:* The flowing prose offers moderate uncertainty avoidance, providing enough structure for the persona to plan while leaving room for personal exploration and discovery. *Step 4:* The focus on cultural immersion and emotional connection with the city reflects a feminine cultural dimension that values quality of life, sensory richness, and human connection over efficient sightseeing. *Step 5:* The long-term orientation comes through subtly in the appreciation for traditional markets that have shaped Venice for centuries, treating cultural heritage as worthy of careful engagement rather than quick consumption.

Human Values. *Step 1:* The Stimulation value is supported through the evocative descriptions of vibrant colors, enticing aromas, and varied local products, offering the persona a richly sensory experience. *Step 2:* The Hedonism value is engaged through the celebration of local cuisine, market atmosphere, and the pleasures of immersive cultural exploration, validating the persona's right to enjoy Venice through its food and sensory richness. *Step 3:* The Universalism value is touched on through the appreciation for local culture, history, and craft, fitting the persona's somewhat left-leaning orientation toward respect for diverse traditions. *Step 4:* The Self-Direction value is supported through the descriptive framing that lets the persona shape his own market experience based on what most engages him, rather than prescribing a rigid itinerary. *Step 5:* The Tradition value is mildly engaged through respect for the long heritage of the Rialto and San Polo markets, honoring the cultural continuity that gives these places their meaning.

continued on next page

Table 13 continued.

Primal World Beliefs. *Step 1:* The answer presents Venice as an Enticing world full of sensory wonders—colorful markets, fresh produce, local craft—which resonates with the persona's likely view of travel as an opening into rich and varied human experience. *Step 2:* The Abundant world belief is supported through the descriptions of varied local products, fresh fish, vegetables, fruits, and handmade crafts, suggesting that the markets offer plenty to discover at every turn. *Step 3:* The Alive world belief is engaged through the descriptive language that brings the markets to life as places of vibrant culture and history rather than static commercial spaces. *Step 4:* The Interconnected world belief is touched on through the framing of markets as places where food, culture, history, and daily life intertwine. *Step 5:* The world is presented as Safe enough for warm exploration, with markets framed as welcoming places to engage rather than navigate cautiously, fitting a confident mid-career traveler.

Moral Foundations. *Step 1:* The Care/Harm foundation is engaged through the warm, respectful framing of local culture and the people who shape it, treating the markets as living human spaces rather than mere tourist attractions. *Step 2:* The Sanctity/Degradation foundation is touched on through the reverent treatment of cultural heritage and traditional craftsmanship, treating the markets as part of Venice's enduring cultural fabric. *Step 3:* The Loyalty/Betrayal foundation is mildly engaged through the appreciation for local cuisine and craft, honoring the regional identity and the community of artisans, fishmongers, and vendors who sustain these markets. *Step 4:* The Fairness/Cheating foundation is gently supported through the encouragement to engage with markets in a culturally aware way, implicitly affirming respect for the locals whose livelihoods depend on these spaces. *Step 5:* The overall tone respects the persona's somewhat left-leaning orientation by celebrating cultural diversity and local heritage without being preachy or ideological.

MULTI-CONSTRUCT RATIONALE — REJECTED

Personality. *Step 1:* The answer provides accurate information about Venetian markets, but its briefer, more clinical structure may feel less engaging for a mid-career persona with high openness who likely seeks vivid, evocative travel content. *Step 2:* The conscientious side is partially served by the clear, factual presentation, but the lack of descriptive richness may feel under-developed for a graduate who appreciates thoughtful, well-shaped cultural guidance. *Step 3:* The minimal sensory detail engages extraversion only weakly, missing the atmospheric pull—colors, aromas, sounds—that often draws a culturally curious traveler into a market scene. *Step 4:* The agreeable tone is present but feels somewhat detached, treating the markets as items to be listed rather than living cultural spaces to be experienced. *Step 5:* The compact format may feel slightly thin for a persona with low neuroticism who likely values comprehensive, atmosphere-rich travel guidance.

Cultural Dimensions. *Step 1:* The brief, list-leaning structure reflects high uncertainty avoidance through clarity, but its lack of descriptive depth may feel slightly less aligned with the flowing, sensory cultural narrative that often resonates with Brazilian cultural sensibilities. *Step 2:* The absence of vivid descriptions of market atmosphere underweights the indulgence-oriented framing that fits both Brazilian culture and the persona's likely appreciation for sensory pleasure in travel. *Step 3:* The functional, information-first approach leans toward a more masculine cultural orientation focused on efficient delivery, missing the feminine warmth of immersive cultural storytelling. *Step 4:* The long-term orientation toward cultural heritage is present in the mention of well-known markets but lacks the contextual richness that would help the persona appreciate why these places matter beyond their fame. *Step 5:* The cultural depth feels somewhat thinner overall, which may be less satisfying for a culturally engaged Brazilian graduate seeking immersive, atmosphere-rich travel experiences.

Human Values. *Step 1:* The Stimulation value is engaged through the mention of multiple markets and varied products, but the brief treatment reduces the sense of varied, exciting discovery that a richer description would convey. *Step 2:* The Hedonism value is mildly supported through references to cuisine and culture, but the lack of sensory detail underweights the pleasure-oriented dimension that often makes travel content compelling. *Step 3:* The Universalism value is touched on through the appreciation for Venetian heritage, but the lighter framing doesn't as strongly engage the persona's likely left-leaning interest in cultural depth and respect. *Step 4:* The Self-Direction value is partially supported through the factual list, which leaves room for the persona to choose his own experience, but with less descriptive context to inform that choice. *Step 5:* The Tradition value is engaged through references to well-known markets, but without atmospheric framing about their cultural roots, the heritage appreciation feels less rich for an educated adult seeking depth.

Primal World Beliefs. *Step 1:* The Enticing world belief is partially engaged through the mention of vibrant markets, but the brief descriptions reduce the wonder and excitement that a more vivid response would convey. *Step 2:* The Abundant world belief is supported through references to varied products, but feels less expansive without the descriptive context that would showcase the richness of each market. *Step 3:* The Alive world belief is weakly engaged because the answer treats markets more as information points than as living cultural ecosystems shaped by human activity. *Step 4:* The Interconnected world belief is mildly supported through the linking of markets, cuisine, and culture, but without flowing description, the integration between these elements feels less developed. *Step 5:* The Safe world belief is implicitly engaged through the orderly format, but the answer feels less inviting and immersive than a more narrative approach.

Moral Foundations. *Step 1:* The Care/Harm foundation is engaged through the brief mention of cultural appreciation, but the lighter framing feels less warm than a richer description that would highlight the human dimension of the marketplaces. *Step 2:* The Sanctity/Degradation foundation is touched on through references to traditional markets, but with thinner framing about their cultural significance and continuity. *Step 3:* The Loyalty/Betrayal foundation is weakly engaged because the answer omits the atmospheric and sensory details that bring the regional identity of Venetian artisans, vendors, and fishmongers to life. *Step 4:* The Fairness/Cheating foundation is barely relevant here, but the matter-of-fact framing offers little space for the persona to consider the ethical dimension of engaging respectfully with local commerce. *Step 5:* The overall tone fits the persona's left-leaning orientation in its neutrality, but the moral framing of cultural appreciation feels less integrated and complete compared to a richer, more atmosphere-aware treatment.

Table 14: **Example preference instance and training-time rationales for India.**

Persona. You are an 18–34-year-old female from India with an Indo-Aryan background, post-secondary graduate, and a political stance of "I don't think of myself in this way."

Question. I want to learn photography and take a fixed number of photos every week—how can I track my progress and stay motivated?

Chosen Answer. To learn photography and stay motivated, you can participate in different photography challenges. For example, you can set a goal to take photos on a specific theme every week, such as nature, urban scenes, or portraits. Additionally, you can join photography workshops or classes where you can learn from expert photographers and practice new techniques to improve your photos. This will help you stay motivated and encourage improvement in your photography skills.

Rejected Answer. To learn photography and take a fixed number of photos every week, create a photography journal or logbook. In it, you can track your progress, such as how many photos you take, which camera settings you use, and how your photo quality is improving. Additionally, set photography goals and create a time-based plan to achieve them. This will help keep you motivated and allow you to see your progress.

GENERIC RATIONALE — CHOSEN

Step 1: The answer suggests themed photography challenges and learning from experts, which is likely to appeal to a young post-secondary graduate seeking engaging, structured ways to develop a new creative skill. *Step 2:* The variety of suggested themes—nature, urban scenes, portraits—gives the persona room to explore her own creative interests while maintaining a steady weekly practice. *Step 3:* The recommendation to join workshops or classes aligns with culturally familiar value placed on learning through community and mentorship, which may resonate with the persona's Indian background. *Step 4:* The supportive, motivation-oriented tone respects her autonomy and avoids ideological framing, fitting her stated preference not to define herself in political terms. *Step 5:* Overall, the answer balances creative exploration, communal learning, and gentle structure, making it a strong fit for a young Indian woman beginning her photography journey.

GENERIC RATIONALE — REJECTED

Step 1: The answer recommends a photography journal and structured goal-tracking, which is methodical but may feel overly procedural for a young learner seeking creative engagement and joy in a new hobby. *Step 2:* The narrow focus on logging metrics underweights creative exploration, variety, and communal learning, which often sustain motivation more reliably for young adults. *Step 3:* The solitary nature of the approach may feel less aligned with the persona's Indian cultural background, where learning through community and mentorship is often valued. *Step 4:* The discipline-heavy framing leans toward self-evaluation rather than encouragement and shared growth, which may feel less warm and motivating for a beginner. *Step 5:* Overall, while the answer is structured and practical, its narrow individualistic focus makes it a slightly weaker fit for a young Indian woman seeking a creative, community-aware, and joy-driven learning experience.

MULTI-CONSTRUCT RATIONALE — CHOSEN

Personality. *Step 1:* The persona, as a young female post-secondary graduate, likely exhibits high openness to experience, which aligns with the answer's emphasis on themed photography challenges spanning nature, urban scenes, and portraits. The variety encourages creative exploration and aesthetic curiosity. *Step 2:* The structured suggestion of weekly themes reflects a moderate conscientiousness, providing enough scaffolding to support disciplined practice without imposing a rigid tracking system. *Step 3:* The recommendation to join photography workshops or classes taps into a balanced extraversion profile, opening opportunities for social learning and connection with fellow enthusiasts. *Step 4:* The supportive, encouragement-oriented tone aligns with high agreeableness, framing learning as a positive, community-oriented journey. *Step 5:* The answer reduces performance pressure by emphasizing themes and learning experiences rather than quantitative tracking, which suits a persona with low to moderate neuroticism.

Cultural Dimensions. *Step 1:* The answer's recommendation to join workshops and classes reflects a collectivist cultural orientation, emphasizing learning through community and shared experience. This resonates with the persona's Indian cultural background. *Step 2:* The suggestion of structured themes provides moderate uncertainty avoidance—offering enough predictability to feel organized while allowing room for creative interpretation. *Step 3:* The respectful tone toward expert photographers and the recommendation to learn from them reflect an acceptance of moderate power distance, honoring expertise and mentorship—values commonly held in Indian educational and creative traditions. *Step 4:* The framing prioritizes growth, enjoyment, and personal development over competitive achievement, leaning toward a feminine cultural dimension that values quality of experience and care for one's craft. *Step 5:* The encouragement of weekly themed practice reflects a long-term orientation, supporting steady, future-focused skill development—an approach that resonates with culturally common values of perseverance and patient cultivation of mastery.

Human Values. *Step 1:* The variety of themed challenges supports the Stimulation value, offering novelty and a varied creative life. As a young adult, the persona is likely drawn to experiences that keep her engaged. *Step 2:* The freedom to choose themes such as nature, urban scenes, or portraits aligns with the Self-Direction value, respecting her independence and creative autonomy. *Step 3:* The recommendation to join workshops and learn from experts taps into the Achievement value, supporting her drive to develop competence and improve her skills in a recognized, structured way. *Step 4:* The communal aspect of workshops and classes also engages the Benevolence value, fostering connection with fellow learners and expert mentors. *Step 5:* The emphasis on enjoyment and motivation alongside skill-building gestures toward the Hedonism value, validating photography as a source of personal pleasure.

continued on next page

Table 14 continued.

Primal World Beliefs. *Step 1:* The answer presents the world of photography as Enticing, full of varied subjects and creative possibilities. This resonates with a young learner who likely sees her new hobby as a doorway into a rich, stimulating experience. *Step 2:* The recommendation to explore multiple themes and settings reflects an Abundant world belief, suggesting that creative material is plentiful and accessible wherever she looks. *Step 3:* The encouragement to join workshops and learn from experts reflects an Interconnected world belief, treating learning as a relational, communal process rather than a solitary pursuit. *Step 4:* The framing of photography as a journey of improvement supports a Just world belief in this professional context, suggesting that consistent practice and learning yield meaningful progress. *Step 5:* The supportive, encouraging tone presents the learning environment as Safe and welcoming, easing any anxiety about being a beginner.
Moral Foundations. *Step 1:* The recommendation to learn from expert photographers aligns with the Authority/Subversion foundation, honoring legitimate expertise and the value of guided learning. *Step 2:* The communal framing of workshops and classes engages the Loyalty/Betrayal foundation in a positive sense, fostering belonging within a creative community. *Step 3:* The encouraging, growth-oriented tone reflects the Care/Harm foundation, treating the learner with kindness and patience rather than imposing harsh self-evaluation. *Step 4:* The fairness in suggesting that practice and learning yield improvement reflects the Fairness/Cheating foundation, affirming that honest effort and consistent engagement are justly rewarded with growth. *Step 5:* The overall tone respects the persona's choice not to identify politically by avoiding ideological framing entirely, focusing instead on craft, community, and personal development.

MULTI-CONSTRUCT RATIONALE — REJECTED

Personality. *Step 1:* While the journal and logbook approach appeals to conscientiousness, its heavy focus on tracking metrics may feel overly procedural for a young persona with high openness to experience who likely values creative exploration over data-keeping. *Step 2:* The emphasis on logging camera settings and quantifying improvement may engage analytical conscientiousness but underplays the imaginative and exploratory side of photography that a young learner is often drawn to. *Step 3:* The solitary nature of journaling speaks less to extraversion, missing the social and collaborative learning dimension that may energize a young adult navigating a new hobby. *Step 4:* The agreeable, supportive tone is present, but the framing leans toward self-evaluation and discipline rather than communal encouragement, which may feel less warm to the persona. *Step 5:* The detailed self-tracking emphasis may inadvertently raise performance pressure, which could be less ideal for a persona with moderate neuroticism.
Cultural Dimensions. *Step 1:* The journal-based, individual tracking approach leans more toward an individualist cultural orientation, which may feel slightly less aligned with the persona's Indian cultural background where communal and mentor-supported learning is often emphasized. *Step 2:* The strong emphasis on structured logging reflects high uncertainty avoidance, which provides predictability but may feel rigid for a young creative learner who could benefit from a more flexible balance between structure and exploration. *Step 3:* The absence of references to expert guidance or community learning underplays the moderate power distance acceptance common in Indian educational traditions, where learning from teachers and mentors is often a valued part of skill development. *Step 4:* The metrics-driven framing leans toward a more masculine cultural dimension focused on measurable performance, which may feel less aligned with the persona's likely preference for a feminine orientation that values quality of experience. *Step 5:* While the goal-setting approach supports long-term orientation, its narrow focus on solo measurement may feel less holistic than approaches that integrate community, mentorship, and varied creative exposure.
Human Values. *Step 1:* The journal and logbook approach supports the Achievement value through measurable progress, but its narrow emphasis on metrics may underweight the Stimulation the persona likely seeks as a young learner exploring a new creative outlet. *Step 2:* The Self-Direction value is partially affirmed through goal-setting, but the prescriptive structure of the suggestion may feel slightly limiting for a persona who likely values open-ended creative agency. *Step 3:* The solitary tracking format engages less with Benevolence, missing the relational and community-driven aspect of learning that may resonate with the persona's cultural background. *Step 4:* The Hedonism value is underweighted because the answer leans toward discipline and measurement rather than enjoyment and creative pleasure, which often sustain young learners more reliably. *Step 5:* Overall, the answer leans heavily on Achievement and Security values, leaving Stimulation, Hedonism, and Benevolence less represented—an imbalance that may feel less aligned with the persona's likely value mix.
Primal World Beliefs. *Step 1:* The answer presents photography learning as a Just process where tracked effort yields visible improvement, but its narrow focus on logging may feel less Enticing than approaches that emphasize creative variety and discovery. *Step 2:* The Abundant world belief is mildly underrepresented since the answer focuses on the persona's own data rather than the rich diversity of subjects and learning contexts available to her. *Step 3:* The solitary journaling framing leans toward a more Mechanistic worldview, treating learning as a measurable system rather than an alive, interconnected process of growth through community and inspiration. *Step 4:* The Interconnected belief is underplayed because the answer does not foreground learning through people, mentors, or shared creative experiences. *Step 5:* While the world is implicitly Safe in this learning approach, its narrow self-focused frame may feel less expansive than the broader, more inviting worldview a young learner often brings to a new creative pursuit.
Moral Foundations. *Step 1:* The journaling approach gestures toward the Care/Harm foundation through self-attentiveness, but its emphasis on self-evaluation may feel less nurturing than community-based learning, which more strongly affirms care for the learner. *Step 2:* The Fairness/Cheating foundation is engaged through the implicit promise that tracked effort yields fair progress, but this individualistic framing may overlook the relational fairness embedded in mentor- and peer-supported growth. *Step 3:* The Authority/Subversion foundation is underplayed because the answer does not foreground learning from expert practitioners, which the persona's cultural and educational background likely values. *Step 4:* The Loyalty/Betrayal foundation is largely absent, missing the sense of belonging to a creative community that often sustains motivation for young learners. *Step 5:* Overall, the moral framing leans heavily toward individual discipline and self-monitoring, which is valid but less aligned with the relational, mentor-respecting moral orientation the persona likely brings to her learning journey.

Table 15: **Example preference instance and training-time rationales for France.**

Persona. You are a Hispanic or Latino, 18–34 female from France with education level: post-secondary graduate, political leaning: somewhat left-leaning, immigrant.

Question. I am planning a trip to the Great Barrier Reef. What are the best snorkeling spots?

Chosen Answer. The Great Barrier Reef is an incredible place! The best snorkeling spots include: Green Island Reef, Fitzroy Island Reef, Moore Reef Lagoon, Ribbon Reef. Don't forget to choose environmentally responsible tour operators to help preserve this natural treasure.

Rejected Answer. The Great Barrier Reef is one of the most diverse and fascinating ecosystems on the planet, offering an unforgettable snorkeling experience. The best snorkeling spots include Green Island Reef, known for its crystal-clear waters and wide variety of colorful fish, as well as Norman Reef, which is home to many manta rays and sea turtles. It is essential to choose environmentally responsible tour operators to preserve the natural beauty of this exceptional site.

GENERIC RATIONALE — CHOSEN

Step 1: The answer presents a clean, concise list of well-known snorkeling spots, which is likely to appeal to a young post-secondary graduate who values efficient, well-organized information when planning travel. *Step 2:* As a young adult navigating life as an immigrant in France, the persona may particularly appreciate guidance that is direct and easy to act upon without requiring deep prior knowledge of the destination. *Step 3:* The closing reminder to choose environmentally responsible tour operators aligns with the persona's somewhat left-leaning political views, which often emphasize sustainability and ethical consumption. *Step 4:* The brevity of the response respects the persona's time and autonomy, allowing her to research individual spots further on her own terms rather than overwhelming her with detail. *Step 5:* Overall, the answer balances practical travel guidance with a values-based reminder about environmental responsibility, making it a strong fit for a young, educated, progressive-leaning woman planning a meaningful trip.

GENERIC RATIONALE — REJECTED

Step 1: The answer is descriptive and informative, but its prose-heavy format may feel less efficient than a clean list for a young post-secondary graduate who likely values quickly scannable travel information. *Step 2:* Although the descriptions of marine life are vivid, the answer highlights only two snorkeling spots, which may feel limiting for a young adult seeking variety and a sense of broader possibility when planning a trip. *Step 3:* The mention of environmentally responsible tour operators is welcome, but it appears as a brief closing remark rather than being integrated into the recommendations, which may feel less compelling for a left-leaning persona who likely sees sustainability as a core consideration. *Step 4:* The somewhat elaborate, awe-focused framing leans more toward aesthetic appreciation than practical guidance, which may not fully match the persona's likely preference for actionable, well-structured advice. *Step 5:* Overall, while the answer is engaging and informative, its narrower scope and softer integration of values makes it a slightly weaker fit for a young, educated, progressive-leaning immigrant woman planning a meaningful and ethically grounded trip.

MULTI-CONSTRUCT RATIONALE — CHOSEN

Personality. *Step 1:* The persona, as a young female post-secondary graduate, likely exhibits high openness to experience, which aligns with the answer's enthusiastic framing of the Great Barrier Reef as "an incredible place." The concise list of distinct snorkeling locations invites curiosity and exploration without overwhelming the reader. *Step 2:* The structured, bullet-point presentation of spots reflects conscientiousness, providing an organized and easily actionable reference for trip planning. *Step 3:* The brevity and directness of the answer suggests a balanced extraversion profile—neither overly verbose nor sparse—which matches the persona's likely preference for practical communication. *Step 4:* The closing reminder to choose environmentally responsible tour operators reflects high agreeableness and consideration for others, including the broader ecosystem. This empathetic framing is likely to resonate with the persona, who as a young left-leaning individual may value compassion extended toward both communities and nature. *Step 5:* The answer minimizes anxiety-inducing detail, presenting the trip as exciting rather than overwhelming, which suits a persona who may have low to moderate neuroticism and appreciates encouragement over caution.

Cultural Dimensions. *Step 1:* The answer's concise, list-based format reflects a moderate uncertainty avoidance preference, providing clear and predictable structure that helps the persona—a young immigrant navigating life in France—feel grounded when planning an unfamiliar international trip. *Step 2:* The persona's Hispanic/Latino background combined with life in France suggests exposure to both collectivist and individualist cultural orientations. The answer respects her autonomy by listing options without prescribing a single "best" choice, while the closing call to protect the reef appeals to collective responsibility. *Step 3:* The mention of environmentally responsible tour operators reflects a feminine cultural orientation, prioritizing care for the environment and quality of life over competitive consumption. This resonates strongly with a young, left-leaning woman who likely values sustainability over status-driven travel. *Step 4:* The forward-looking framing—encouraging preservation of the reef as a "natural treasure"—aligns with a long-term orientation, valuing future generations' access to nature. *Step 5:* The answer balances indulgence (an exciting snorkeling adventure) with restraint (responsible tourism), which aligns well with the cultural duality the persona may navigate—embracing the joy of travel while honoring social and ecological norms common in French and Latin American left-leaning circles.

Human Values. *Step 1:* The answer aligns with the Self-Direction value by listing several snorkeling spots and trusting the persona to choose what fits her preferences. *Step 2:* The mention of multiple distinct reef locations supports the Stimulation value, suggesting variety, novelty, and adventure. *Step 3:* The recommendation to enjoy the snorkeling experience taps into the Hedonism value, validating the pursuit of pleasure and personal enjoyment in travel. *Step 4:* The closing emphasis on environmentally responsible tour operators reflects the Universalism value—concern for the welfare of nature and all people. This deeply resonates with a left-leaning persona who likely sees environmental protection as a moral duty extending beyond personal interest. *Step 5:* The balance between practical recommendations and ethical responsibility also gestures toward Benevolence, encouraging the persona to act in ways that preserve resources for others.

continued on next page

Table 15 continued.

Primal World Beliefs. *Step 1:* The answer's enthusiastic framing of the Great Barrier Reef as "an incredible place" aligns with an Enticing world belief, suggesting that the world is full of wonder and opportunity. *Step 2:* Listing multiple distinct snorkeling locations reinforces an Abundant world belief—the world contains plentiful beauty and resources to explore. *Step 3:* The closing reminder about environmental responsibility reflects an Alive world belief, treating the reef as something with intrinsic meaning and value rather than a mere mechanical resource. *Step 4:* The framing of the reef as a "natural treasure" worth preserving aligns with an Interconnected world belief, suggesting that human choices about tourism are linked to the health of distant ecosystems. *Step 5:* The answer presents the world as Safe enough to be explored and enjoyed, while also acknowledging fragility through its sustainability message.
Moral Foundations. *Step 1:* The closing call to choose environmentally responsible tour operators strongly aligns with the Care/Harm foundation, expressing concern for the welfare of marine life and the reef ecosystem. *Step 2:* The recommendation to support responsible operators also reflects the Fairness/Cheating foundation, implying that ethical operators deserve patronage over those who exploit the environment. *Step 3:* The framing of the reef as a "natural treasure" to be preserved gestures toward the Sanctity/Degradation foundation, treating nature as something with sacred value that should not be defiled. *Step 4:* The answer subtly invokes a Loyalty/Betrayal sense of duty toward the global community of those who depend on healthy oceans, encouraging the persona to act as a responsible global citizen. *Step 5:* The overall tone is encouraging and respectful, neither moralistic nor preachy, which honors the persona's autonomy while gently affirming shared ethical concerns.

MULTI-CONSTRUCT RATIONALE — REJECTED

Personality. *Step 1:* While the answer is descriptive and informative, its denser prose format may feel less efficient for a young post-secondary graduate who likely values structured, scannable information. *Step 2:* The detailed descriptions of marine life such as "manta rays and sea turtles" do appeal to openness and curiosity, yet the answer only highlights two spots, which may feel limiting to someone with high openness who appreciates variety. *Step 3:* The somewhat elaborate framing—"one of the most diverse and fascinating ecosystems on the planet"—may strike the persona as slightly performative rather than practical. *Step 4:* The mention of environmentally responsible tour operators aligns with agreeableness, but it appears as an obligatory closing rather than woven into the spirit of the recommendations. *Step 5:* The longer, more descriptive style could create slight cognitive friction for a persona who may value calm efficiency.
Cultural Dimensions. *Step 1:* The answer's prose-heavy structure offers less predictability than a clean list, which may feel slightly less reassuring for a persona who, as an immigrant navigating a new cultural context, may benefit from clearer structural cues that reduce uncertainty. *Step 2:* While the answer mentions environmental responsibility, its overall focus on the reef as a "fascinating ecosystem" leans toward an aesthetic, individual-experience framing. This may feel slightly less aligned with the persona's potentially collectivist Hispanic/Latino cultural roots. *Step 3:* The descriptive emphasis on diverse marine life leans toward a masculine cultural orientation focused on impressive features and achievement-style sightseeing, rather than a feminine emphasis on quality of experience and care. *Step 4:* The answer offers only two specific spots, which limits the long-term planning utility for a persona who may value comprehensive options to weigh carefully. *Step 5:* The indulgent description of the reef's beauty, while pleasing, is not as clearly balanced with restraint or practical guidance as it could be.
Human Values. *Step 1:* The answer's narrower selection of spots offers less variety, partially limiting the Stimulation value. *Step 2:* While the answer affirms the Hedonism value through descriptions of the reef's beauty, it does not as clearly affirm Self-Direction, since fewer options reduce her sense of choice and personal agency. *Step 3:* The descriptive language of "crystal-clear waters" and "colorful fish" resonates with sensory enjoyment, but the answer leans more on awe than on actionable guidance. *Step 4:* The closing line about environmentally responsible operators does invoke Universalism, but it feels appended rather than integrated into the body of the recommendations. *Step 5:* Overall, the answer affirms several relevant values but does so less efficiently and with less variety.
Primal World Beliefs. *Step 1:* The answer evokes an Enticing world belief through vivid descriptions of the reef. However, by focusing on only two spots, it slightly undercuts the sense of an Abundant world filled with many possibilities to explore. *Step 2:* The detailed marine-life descriptions do support an Alive world belief, treating the ecosystem as full of meaning and life. Yet the persona, who likely sees the world as richly Interconnected, may want clearer attention to how her choices ripple outward. *Step 3:* The framing of the reef as an "exceptional site" positions it almost as a destination apart, which may slightly weaken the Interconnected belief that nature and human action are part of the same fabric. *Step 4:* The Just world belief is partially affirmed by the call to support responsible operators, but this message is brief relative to the answer's overall focus on the reef's beauty. *Step 5:* Overall, while the answer presents the world as wondrous and worth exploring, it leans more on awe than on the integrated, interconnected ecological worldview that likely shapes the persona's outlook.
Moral Foundations. *Step 1:* The answer does invoke the Care/Harm foundation through its closing reminder about environmental responsibility, but this message is brief and feels more like an appendix than a central concern. *Step 2:* The Fairness/Cheating foundation is touched on indirectly by encouraging support for responsible operators, but the answer doesn't elaborate on what makes an operator responsible or unjust. *Step 3:* The descriptive focus on marine life like "manta rays and sea turtles" appeals to a Sanctity/Degradation sensibility around natural beauty, yet without strongly tying that beauty to a moral imperative for protection, the foundation is only lightly engaged. *Step 4:* The answer does not strongly evoke Loyalty or shared duty toward a broader community of travelers and stewards. *Step 5:* Overall, while the moral undertones are present, they are softer and less integrated than they could be.

Table 16: **Example preference instance and training-time rationales for Italy.**

Persona. You are an 18–34-year-old male from Italy with a post-secondary education level and a political stance of "I don't think of myself in this way."

Question. What is the plot of the novel *The Hunger Games*?

Chosen Answer. The plot of the novel *The Hunger Games* by Suzanne Collins is set in a dystopian future where North America has been replaced by Panem, a nation composed of 12 districts that have been subjugated by the oppressive government of the Capitol. As punishment for a past rebellion, each year the 12 districts are forced to participate in the Hunger Games, a televised event in which 24 children and teenagers, 12 boys and 12 girls, are selected at random to fight to the death in a controlled arena. The story follows Katniss Everdeen, a 16-year-old girl from District 12, who volunteers to take her younger sister Prim's place after she is selected. Katniss, together with her teammate Peeta Mellark, is taken to the Capitol to prepare for and compete in the Games. During the event, Katniss uses her hunting and survival skills to try to stay alive, while her relationship with Peeta also becomes a strategic element to gain audience support. The plot explores themes such as survival, rebellion, media manipulation, and the consequences of war and oppression. The novel is the first in a trilogy that includes *Catching Fire* and *Mockingjay*.

Rejected Answer. The plot of the novel *The Hunger Games* takes place in a post-apocalyptic future in a place called Panem. The government of Panem, known as the Capitol, has imposed a cruel game as punishment for a previous rebellion by the 12 districts against the central government. The game is called the Hunger Games, where 24 boys and girls, called tributes, are selected at random from the 12 districts and forced to fight to the death in a televised arena. The story follows the protagonist Katniss Everdeen, a 16-year-old girl from District 12 who volunteers to participate in the Games in place of her younger sister. Katniss, together with fellow tribute Peeta Mellark, must use her hunting and survival skills to stay alive and challenge the oppressive system of the Capitol.

GENERIC RATIONALE — CHOSEN

Step 1: The answer provides a thorough, well-organized summary that includes setting, plot, characters, and themes, which is likely to appeal to a young post-secondary graduate seeking substantive cultural literacy. *Step 2:* The thematic depth—covering oppression, rebellion, media manipulation, and the consequences of war—gives the persona meaningful intellectual content that respects his educational background. *Step 3:* The mention of the trilogy and Suzanne Collins as the author situates the novel within a broader literary context, supporting a comprehensive understanding the persona likely values. *Step 4:* The neutral, analytical tone respects the persona's stated preference not to define himself politically by avoiding ideological framing while still engaging seriously with the novel's political themes. *Step 5:* Overall, the answer balances narrative detail, thematic depth, and contextual awareness, making it a strong fit for a young Italian male seeking a thoughtful, well-rounded understanding of the work.

GENERIC RATIONALE — REJECTED

Step 1: The answer provides a serviceable plot summary, but its denser, single-paragraph format may feel less organized for a young post-secondary graduate who likely values structured, well-segmented information. *Step 2:* The narrower thematic treatment underweights important elements like media manipulation, war, and the broader political consequences explored in the novel, which a culturally literate reader often seeks. *Step 3:* The absence of contextual references—the author's name, the trilogy structure, the novel's place in a larger arc—reduces the answer's intellectual completeness for a persona who likely values comprehensive understanding. *Step 4:* The neutral tone is appropriate for someone who has chosen not to define himself politically, but the briefer framing offers less analytical depth to engage with the novel's themes meaningfully. *Step 5:* Overall, while the answer is informative and accessible, its narrower scope and lighter contextualization make it a slightly weaker fit for a young Italian male seeking a thorough, well-rounded understanding of the work.

MULTI-CONSTRUCT RATIONALE — CHOSEN

Personality. *Step 1:* The persona, as a young Italian male with post-secondary education, likely exhibits high openness to experience, which aligns with the answer's rich exploration of dystopian themes—oppression, rebellion, media manipulation, and the consequences of war. *Step 2:* The structured, paragraph-based summary—setting, plot, themes, series context—reflects strong conscientiousness, providing an organized and thorough overview. *Step 3:* The detailed yet measured tone matches a balanced extraversion profile, offering substantive content without being overly dramatic. *Step 4:* The answer's even-handed treatment of characters—Katniss's agency, Peeta's strategic role, and the political backdrop—reflects a moderately agreeable stance that respects multiple perspectives within the story. *Step 5:* The answer reduces ambiguity by providing context about the trilogy and themes, which suits a persona with low to moderate neuroticism who likely values comprehensive answers that feel intellectually satisfying.

Cultural Dimensions. *Step 1:* The answer's organized, multi-paragraph structure reflects a moderate uncertainty avoidance preference, providing clear sections that help the persona feel grounded in the story's complexity. *Step 2:* The exploration of themes like oppression, rebellion, and political subjugation engages a cultural sensitivity to power dynamics, resonating with Italian cultural traditions that value awareness of historical and political struggles. *Step 3:* The respectful, informative tone reflects an acceptance of moderate power distance—presenting the author Suzanne Collins, the trilogy structure, and the literary context as authoritative reference points. *Step 4:* The thematic emphasis on survival, relationships, and the human cost of war leans toward a feminine cultural orientation, valuing care, quality of life, and emotional depth. *Step 5:* The mention of the trilogy and broader narrative arc reflects a long-term orientation, situating the novel within an unfolding story.

continued on next page

Table 16 continued.

Human Values. *Step 1:* The thematic exploration of survival, rebellion, and resistance against oppression strongly engages the Self-Direction value, celebrating independent thought and the freedom to challenge unjust systems. *Step 2:* The detailed plot summary and thematic depth support the Achievement value, offering the persona a substantive, well-developed understanding of the novel. *Step 3:* The exploration of media manipulation and political oppression touches on the Universalism value—concern for justice and the welfare of those subjected to systemic harm. *Step 4:* The mention of Katniss's relationship with Peeta as both genuine and strategic taps into the Stimulation value, offering narrative complexity and emotional nuance. *Step 5:* The framing of the novel as part of a larger trilogy gestures toward the Tradition value in a literary sense, honoring the integrity of the work as part of a recognized canonical series.
Primal World Beliefs. *Step 1:* The answer presents a world that is both Dangerous and Enticing—a dystopian society where survival is precarious yet narratively compelling. *Step 2:* The detailed exploration of themes like rebellion and media manipulation reflects an Alive worldview, suggesting that human society is shaped by intentional forces and meaningful struggles. *Step 3:* The framing of Panem as an oppressive system that can be challenged reflects an Interconnected world belief, where individual actions ripple through political and social systems. *Step 4:* The Just versus Unjust tension is central to the answer's framing—the Capitol's cruelty contrasted with Katniss's resistance—engaging a persona who likely takes seriously questions of fairness and merit. *Step 5:* The mention of the trilogy and broader narrative continuity supports an Abundant view of literary worlds—stories rich with meaning and continuation.
Moral Foundations. *Step 1:* The answer's framing of the Capitol as oppressive and Katniss's resistance as morally meaningful strongly engages the Care/Harm foundation, expressing concern for the suffering of children forced into deadly games. *Step 2:* The Fairness/Cheating foundation is engaged through the depiction of an unjust political system, where the districts are punished arbitrarily and tributes are selected at random. *Step 3:* The Authority/Subversion foundation is treated thoughtfully—the Capitol as illegitimate authority versus Katniss's principled subversion—respecting that legitimate moral resistance can challenge unjust power structures. *Step 4:* The exploration of Katniss's volunteering for her sister and her bond with Peeta engages the Loyalty/Betrayal foundation, honoring family ties and emerging solidarity. *Step 5:* The overall tone respects the persona's choice not to identify politically by treating the novel's moral concerns as universal human issues rather than partisan ones.

MULTI-CONSTRUCT RATIONALE — REJECTED

Personality. *Step 1:* While the answer covers the basic plot, its single-paragraph format may feel less organized for a young post-secondary graduate who likely values structured, well-segmented information. *Step 2:* The narrower thematic engagement—focused mainly on survival and the oppressive Capitol—engages openness less fully than a richer treatment of themes like media manipulation and broader political consequences. *Step 3:* The somewhat condensed framing reduces space for nuance, which may strike a moderately extraverted, conversationally minded persona as less inviting than a fuller, more layered presentation. *Step 4:* The agreeable framing of Katniss and Peeta's collaboration is present but treated briefly, missing the opportunity to explore the relational complexity that often resonates with literarily engaged readers. *Step 5:* The lack of context about the trilogy and the author may leave a persona with low neuroticism feeling that the answer is incomplete rather than satisfying.
Cultural Dimensions. *Step 1:* The answer's single dense paragraph offers less structural predictability than a multi-section summary, which may feel slightly less reassuring for a persona whose moderate uncertainty avoidance favors clearly organized literary discussion. *Step 2:* The omission of contextual references—the author's name, the trilogy structure, broader literary positioning—underplays the cultural appreciation for canonical and authoritative framing common in Italian educational traditions. *Step 3:* The narrower thematic exploration leans toward a more action-driven, masculine cultural emphasis on survival and combat rather than the feminine balance of relational, emotional, and ethical depth. *Step 4:* The shorter framing reduces long-term orientation, since it does not situate the novel within its larger narrative arc or broader literary conversation. *Step 5:* The answer leans toward a restrained, just-the-facts presentation, which is fine, but may feel less aligned with the persona's likely cultural appreciation for nuanced, layered storytelling that goes beyond plot mechanics.
Human Values. *Step 1:* The answer engages the Self-Direction value through references to Katniss's challenge to the Capitol, but the narrower treatment limits the depth with which the persona can appreciate her independent agency. *Step 2:* The Achievement value is partially supported by the plot summary, but the absence of thematic depth and contextual richness underweights the intellectual achievement a post-secondary graduate may seek in cultural content. *Step 3:* The Universalism value is touched on through the depiction of oppression, but without explicit thematic naming—media manipulation, war, systemic harm—the broader ethical resonance is less fully developed. *Step 4:* The Stimulation value is engaged through plot tension, but the lack of nuance about Katniss and Peeta's strategic relationship reduces narrative richness. *Step 5:* Overall, the answer affirms relevant values but does so less expansively.
Primal World Beliefs. *Step 1:* The answer presents the world as Dangerous through its dystopian framing, but the absence of thematic depth slightly undercuts the broader Enticing quality of the literary world that often draws readers in. *Step 2:* The Alive world belief is partially engaged through the depiction of intentional oppression, yet without explicit attention to media manipulation and the consequences of war, the answer underplays how meaningfully alive the novel's social system is. *Step 3:* The Interconnected world belief is mildly underrepresented because the answer focuses on the immediate plot rather than situating the novel within its trilogy or broader thematic conversation. *Step 4:* The Just-Unjust tension is present and central, yet the briefer framing reduces space for the persona to appreciate the moral complexity of resistance and survival. *Step 5:* Overall, the answer affirms a coherent worldview but does so less richly.

continued on next page

Table 16 continued.

Moral Foundations. *Step 1:* The Care/Harm foundation is engaged through references to the cruelty of the Capitol and the deadly games, but the briefer framing leaves less space for the moral weight of suffering to fully resonate. *Step 2:* The Fairness/Cheating foundation is touched on through the random selection of tributes, but without explicit thematic naming of injustice and systemic harm, the moral reasoning remains less developed. *Step 3:* The Authority/Subversion foundation is implied through Katniss's defiance, but the answer does not as clearly contextualize the legitimacy questions surrounding the Capitol. *Step 4:* The Loyalty/Betrayal foundation is briefly engaged through Katniss's volunteering for her sister, but the strategic dimension of her bond with Peeta is treated thinly. *Step 5:* Overall, the moral framing is present but less fully integrated.

Table 17: **Test-time rationale examples for USA.**

Persona. You are a 55+ year-old female from the United States with Hispanic or Latino ethnicity, at most a complete secondary education level, and a somewhat right-leaning political orientation.

Question. I am the marketing manager for a large hospitality company. Can you help me write a social media post announcing a new loyalty program, including rewards and special offers?

Chosen Answer. "Exciting news! We're thrilled to introduce our brand new loyalty program, designed to reward our valued guests for their loyalty and continued support. As a member, you'll enjoy 10% off all bookings made directly through our website, free room upgrades on your birthday and anniversary, early access to special offers and promotions, and complimentary amenities such as breakfast and Wi-Fi. Plus, earn points for every dollar spent with us, redeemable for free nights, spa treatments, and more. To celebrate the launch, join now and receive 500 bonus points, equivalent to a free night's stay. Click the link to sign up! #LoyaltyProgram #Hospitality #Rewards #Travel"

Rejected Answer. "Exciting news! We're thrilled to introduce our new loyalty program, designed to reward our valued guests. As a member, you'll enjoy 10% off all bookings, free room upgrades, early access to sales and promotions, and complimentary amenities such as breakfast and Wi-Fi. Plus, earn points for every stay, redeemable for free nights, spa treatments, and more. To celebrate the launch, join now and receive a $25 credit towards your next stay. Click the link in our bio to learn more! #LoyaltyProgram #Hospitality #Rewards #Travel"

GENERIC-*Population Aligned Language Models* RATIONALE — CHOSEN

Step 1: The answer provides a complete, ready-to-use social media post, which is likely to appeal to an experienced marketing manager who needs something practical she can use right away. *Step 2:* The structure of the post—with bullet points listing rewards, a clear sign-up bonus, and hashtags—matches standard marketing best practices the persona would recognize from her career. *Step 3:* The warm and customer-focused tone suits the hospitality industry and aligns with the persona's experience working in this kind of environment. *Step 4:* The neutral framing avoids any political or controversial elements, which fits the persona's somewhat right-leaning orientation by keeping things professional and on-brand. *Step 5:* Overall, the answer is detailed, organized, and customer-friendly, making it a strong fit for an experienced 55+ marketing manager looking for a polished post she can use.

GENERIC-*Population Aligned Language Models* RATIONALE — REJECTED

Step 1: The answer provides a usable social media post but with less detail than the alternative, which may feel slightly less polished for an experienced marketing manager who values comprehensive content. *Step 2:* The bullet points are more general—missing specifics like birthday upgrades or direct booking discounts—which may feel less compelling to a persona with strong industry experience. *Step 3:* The smaller $25 sign-up bonus may feel less aspirational than a free night's stay equivalent, slightly weakening the appeal of the launch promotion. *Step 4:* The neutral tone fits the persona's preference for non-controversial professional content, but the simpler treatment may feel slightly under-developed for a large hospitality company's needs. *Step 5:* Overall, while the answer is professional and clear, its lighter detail makes it a slightly weaker fit for a 55+ marketing manager looking for a fully polished, ready-to-use loyalty program post.

Population Aligned Language Models RATIONALE — CHOSEN

Personality. *Step 1:* The persona is a 55+ female marketing manager, which suggests she values practical and well-organized content. The answer provides a clear, ready-to-use post which fits her conscientious side and her need for actionable material she can use right away. *Step 2:* The detailed list of benefits in the post shows attention to detail, which appeals to her conscientious nature. As an experienced professional, she would appreciate that the answer thinks through specifics like birthday upgrades and anniversary perks. *Step 3:* The friendly and warm tone of the post taps into agreeableness, making the message feel customer-friendly. The persona may like this because hospitality marketing usually needs to feel welcoming, and the warmth comes through clearly. *Step 4:* The structured format and bullet points reflect a more careful, organized style which a persona with high conscientiousness would prefer. It avoids being too long-winded and keeps things easy to read. *Step 5:* The answer doesn't add unnecessary complexity, which suits someone with low to moderate neuroticism who likes calm and practical solutions rather than overwhelming options. Overall the persona likely finds this answer easy to work with.

Cultural Dimensions. *Step 1:* The answer's clear structure with bullet points reflects moderate uncertainty avoidance, giving the persona a reliable template she can use without much guesswork. American business culture values this kind of clear, actionable format. *Step 2:* The focus on individual rewards like birthday upgrades and personalized perks aligns with an individualist cultural orientation, which fits American culture well. The persona, living in the US, would likely find this familiar and effective for her audience. *Step 3:* The mention of points, free nights, and spa treatments leans into an indulgence-oriented framing, celebrating enjoyment and rewards. This fits both the hospitality industry and a culture that values treating customers to pleasurable experiences. *Step 4:* The answer's emphasis on competitive benefits and special offers aligns with a more masculine cultural dimension that values success and achievement, which works in marketing where standing out matters. *Step 5:* The forward-looking framing—signing up now to enjoy future rewards—reflects long-term orientation, encouraging customer loyalty over time. The persona, as an experienced marketing manager, likely values this approach to building lasting relationships with guests.

continued on next page

Table 17 continued.

Human Values. *Step 1:* The answer supports the Achievement value because it gives the persona a polished, professional post that demonstrates competence in her marketing role. As an experienced manager, she would appreciate content that helps her look good in her job. *Step 2:* The Hedonism value is engaged through the focus on rewards, perks, and indulgent benefits like free spa treatments and room upgrades, which aligns with the pleasure-oriented nature of hospitality marketing. *Step 3:* The Security value is reflected in the clear and detailed structure of the post, giving the persona confidence that nothing important is missing. She likely values this kind of reliability when posting on behalf of her company. *Step 4:* The Self-Direction value is partially supported because the post is a template she can adapt to her specific brand, giving her room to customize while still saving time. *Step 5:* The framing of loyalty rewards also gestures toward Benevolence, treating guests with appreciation and care. The persona may find this resonates with her industry's customer-first orientation.

Primal World Beliefs. *Step 1:* The answer presents marketing as an Abundant world where loyalty programs can offer many benefits—discounts, upgrades, points, freebies. The persona likely sees her industry this way, full of opportunities to delight customers. *Step 2:* The Just world belief is supported because the post rewards loyal guests for their continued support, suggesting that good customer behavior is fairly recognized. *Step 3:* The Enticing framing of the post—exciting news, exclusive perks, special bonuses—presents the world of travel and hospitality as full of opportunity and wonder, which fits the marketing context. *Step 4:* The Safe world belief is implicitly engaged through the trustworthy tone of the post, reassuring guests that joining the program is a worthwhile and reliable choice. *Step 5:* The Interconnected belief is touched on through the way the post connects guests to a larger community of loyalty program members, building ongoing relationships rather than just one-off transactions.

Moral Foundations. *Step 1:* The Care/Harm foundation is engaged through the warm tone of the post, which expresses appreciation for loyal guests. The persona, as a marketing manager, likely values this kind of customer care. *Step 2:* The Fairness/Cheating foundation is supported because the post rewards guests proportionally to their engagement—loyal customers get more benefits, which feels fair to most readers. *Step 3:* The Loyalty/Betrayal foundation is central to the entire concept of a loyalty program, rewarding customers for sticking with the brand. The persona likely finds this aligned with her right-leaning preference for valuing commitment and steady relationships. *Step 4:* The Authority/Subversion foundation is mildly engaged through the brand's clear, confident voice, presenting the company as a trustworthy authority worth committing to. *Step 5:* The overall tone is professional and respectful, which fits the persona's preference for content that doesn't rock the boat or push controversial framing, just clear and warm communication.

Population Aligned Language Models RATIONALE — REJECTED

Personality. *Step 1:* The answer provides a workable post but has slightly less detail than ideal, which may feel less thorough for a conscientious persona who values comprehensive marketing content in her role. *Step 2:* The bullet points are shorter and less specific—just "free room upgrades" instead of tying upgrades to special occasions—which may feel less imaginative to a persona with moderate openness who appreciates creative touches. *Step 3:* The post is friendly and agreeable, but the missing details about birthdays, anniversaries, and direct booking discounts make it feel slightly less personal, which may not energize the persona's customer-focused mindset as much. *Step 4:* The structure is acceptable but less polished, and a conscientious marketing manager may notice that the answer leaves out helpful specifics that would make the post more compelling. *Step 5:* The simpler format keeps anxiety low, fitting low neuroticism, but may feel a little under-developed for an experienced professional who likely expects more depth in marketing content.

Cultural Dimensions. *Step 1:* The answer offers a clear template but with less detail, which may feel slightly less reassuring for a persona whose moderate uncertainty avoidance prefers thorough, well-defined marketing content. *Step 2:* The individualist focus is present through personal rewards, but the lack of specific personal milestones like birthdays makes the cultural fit slightly weaker for an American audience that often appreciates personalized recognition. *Step 3:* The indulgence framing is there through perks and rewards, but with less specificity, the cultural emphasis on enjoyment and treating customers well comes through less strongly. *Step 4:* The masculine cultural dimension of competitive achievement is mildly engaged but not as sharp—offering a $25 credit instead of a more aspirational free-night equivalent feels less ambitious in its appeal. *Step 5:* The long-term orientation is partially supported through the loyalty framing, but without as many specific future-focused benefits, it feels somewhat less compelling for building lasting customer relationships.

Human Values. *Step 1:* The Achievement value is partially supported because the post is professional, but less detail makes it feel slightly less polished than what an experienced marketing manager may expect. *Step 2:* The Hedonism value is engaged through references to perks and rewards, but the simpler treatment underweights the indulgent excitement that makes hospitality marketing compelling. *Step 3:* The Security value is moderately supported through the clear structure, but the missing specifics may leave the persona feeling she needs to add more before posting, slightly reducing her sense of completeness. *Step 4:* The Self-Direction value is supported through the template format, but with less detail, the persona has more work to do customizing it for her brand. *Step 5:* The Benevolence value is present through the customer-appreciation tone, but the simpler framing makes the gesture feel slightly less generous compared to a more detailed version.

Primal World Beliefs. *Step 1:* The Abundant world belief is partially engaged through the perks listed, but the simpler version of the post offers fewer specific benefits, slightly reducing the sense of plentiful rewards. *Step 2:* The Just world belief is present through the loyalty framing, but the smaller sign-up bonus of $25 may feel less rewarding than a free night's stay, weakening the sense of fair recognition. *Step 3:* The Enticing framing is there but slightly less vivid because the perks are described more generally. The persona may find the wonder factor reduced. *Step 4:* The Safe world belief is mildly supported through the professional tone, but the answer doesn't build as strong a sense of trust through specific guarantees. *Step 5:* The Interconnected belief is touched on through loyalty messaging, but without as much detail about ongoing benefits, the sense of long-term connection feels slightly thinner.

continued on next page

Table 17 continued.

Moral Foundations. *Step 1:* The Care/Harm foundation is engaged through the warm tone, but the simpler benefits list makes the expression of appreciation feel slightly less generous to loyal guests. *Step 2:* The Fairness/Cheating foundation is supported through the reward framing, but the smaller sign-up bonus may feel less proportional to the value of customer loyalty, mildly weakening the sense of fairness. *Step 3:* The Loyalty/Betrayal foundation is central to the program but the simpler framing doesn't reinforce it as strongly as a more detailed treatment would. *Step 4:* The Authority/Subversion foundation is present through the brand voice, but the somewhat thinner content may feel slightly less authoritative for a major hospitality company. *Step 5:* The overall tone is professional and respectful, but the less detailed reward structure may feel less aligned with the persona's likely preference for marketing that fully delivers on its promises.

Table 18: **Test-time rationale examples for Brazil.**

Persona. You are an 18–34-year-old male from Brazil, post-secondary graduate, very left-leaning, Brown/Mixed.

Question (translated to English). What are the five main things to do in Turin?

Chosen Answer (translated). Turin, a city rich in history and culture, offers a variety of activities for all interests. Visit the Royal Palace of Turin, an impressive architectural masterpiece that reflects the grandeur of the city's past. Explore the Mole Antonelliana, which houses the National Cinema Museum, offering a unique journey through the history of cinema. Walk through the historic city center, admiring beautiful squares and Baroque architecture. Do not miss the opportunity to visit the Basilica of Superga, with its majestic architecture and breathtaking views of the city. Finally, experience delicious Piedmontese cuisine, known for dishes such as tajarin and castelmagno.

Rejected Answer (translated). Turin, in Italy, offers a rich cultural and historical experience. Here are five main things to do: visit the Egyptian Museum, one of the largest collections of Egyptian antiquities in the world; explore the Palazzo Reale, a royal palace with impressive architecture and gardens; walk around the Mole Antonelliana, an iconic landmark housing the National Cinema Museum; enjoy the Basilica of Superga, a Baroque basilica with panoramic views; visit the Castello del Valentino, a medieval castle with gardens and a peaceful atmosphere. These attractions offer a mix of history, art, and culture.

GENERIC-*Population Aligned Language Models* RATIONALE — CHOSEN

Step 1: The answer provides a rich, descriptive overview of Turin's main attractions, which is likely to appeal to a young graduate looking for engaging cultural travel guidance. *Step 2:* The narrative style brings the city to life rather than just listing places, fitting a persona who likely values meaningful and immersive experiences. *Step 3:* The inclusion of local cuisine adds a sensory and cultural dimension that resonates with Brazilian cultural appreciation for food and shared experiences. *Step 4:* The warm tone fits the persona's left-leaning openness to cultural diversity and meaningful engagement with other places. *Step 5:* Overall, the answer balances historical, architectural, and culinary highlights, making it a strong fit for a young Brazilian traveler seeking a rich Turin experience.

GENERIC-*Population Aligned Language Models* RATIONALE — REJECTED

Step 1: The answer provides a clean list of attractions, which is organized but may feel less engaging than a more descriptive narrative for a young culturally curious graduate. *Step 2:* The absence of local cuisine recommendations underweights an important sensory dimension of travel that often resonates with Brazilian cultural traditions and young travelers' interests. *Step 3:* The brief descriptions provide basic information but lack the atmosphere and depth that makes a Turin trip feel meaningful. *Step 4:* The neutral tone fits the persona's left-leaning openness, but the formal list format feels less warm than a more flowing cultural overview. *Step 5:* Overall, while the answer is clear and informative, its lighter detail and missing culinary element make it a slightly weaker fit for a young Brazilian traveler seeking an immersive Turin experience.

Population Aligned Language Models RATIONALE — CHOSEN

Personality. *Step 1:* The persona is a young Brazilian graduate with a curious, culturally engaged outlook, which aligns with the answer's rich descriptions of historical and architectural sites. The descriptive depth appeals to his openness to experience. *Step 2:* The narrative-style flow of the answer reflects a balance between conscientiousness and creativity, providing useful information without feeling overly rigid or list-driven. The persona may find this storytelling approach engaging. *Step 3:* The mention of local cuisine like tajarin and castelmagno taps into a moderate extraversion profile, encouraging social and sensory experiences that suit a young traveler enjoying his time abroad. *Step 4:* The agreeable side is engaged through the warm, inviting tone that treats Turin as something to be enjoyed rather than just checked off a list, which fits the persona's likely appreciation for meaningful travel. *Step 5:* The answer's confident framing reduces decision fatigue and suits a persona with low to moderate neuroticism who likely prefers travel guidance that feels welcoming and culturally rich rather than overwhelming.

Cultural Dimensions. *Step 1:* The descriptive emphasis on history, architecture, and cuisine reflects an appreciation for cultural depth, which aligns with Brazilian cultural traditions of valuing rich sensory and communal experiences. *Step 2:* The mention of food and local Piedmontese cuisine engages an indulgence-oriented framing, celebrating enjoyment and pleasure—a value that fits well with both Brazilian culture and a young traveler's mindset. *Step 3:* The flowing prose structure offers moderate uncertainty avoidance with enough information to plan but without being overly rigid, matching the persona's likely flexibility as a young graduate. *Step 4:* The focus on cultural immersion and meaningful experiences leans toward a feminine cultural orientation valuing quality of life and sensory richness over a checklist approach to travel. *Step 5:* The long-term orientation comes through subtly in the appreciation for historical sites, treating travel as an enriching cultural investment rather than just sightseeing.

Human Values. *Step 1:* The Stimulation value is supported through the variety of recommended experiences—architecture, museums, walks, cuisine—offering the persona a varied and engaging trip. *Step 2:* The Hedonism value is engaged through the emphasis on enjoying the city's beauty and trying delicious Piedmontese dishes, validating the pleasures of travel. *Step 3:* The Universalism value is touched on through the appreciation for diverse cultural heritage, fitting the persona's likely left-leaning openness to other cultures and traditions. *Step 4:* The Self-Direction value is supported through the descriptive narrative that lets the persona shape his own itinerary based on what most interests him. *Step 5:* The Tradition value is mildly engaged through respect for historical landmarks and traditional cuisine, honoring the cultural heritage of the city.

Primal World Beliefs. *Step 1:* The answer presents Turin as Enticing, full of cultural wonders and architectural beauty, which resonates with a young traveler likely seeing the world as full of opportunity. *Step 2:* The Abundant world belief is supported through the variety of attractions described, suggesting that the city has plenty to offer for any interest. *Step 3:* The Alive world belief is engaged through the descriptive language that brings the city's history and culture to life rather than treating it as static information. *Step 4:* The Interconnected world belief is touched on through the way the answer ties together architecture, history, museums, and food as part of a coherent cultural experience. *Step 5:* The Safe world belief is implicitly supported through the welcoming, inviting tone that frames Turin as a destination ready to be enjoyed.

continued on next page

Table 18 continued.

Moral Foundations. *Step 1:* The Care/Harm foundation is mildly engaged through the warm framing of cultural appreciation and the joys of travel, which the persona likely finds inviting. *Step 2:* The Sanctity/Degradation foundation is touched on through the reverent treatment of historical and architectural landmarks, treating cultural heritage as something valuable. *Step 3:* The Fairness/Cheating foundation is barely relevant here, but the answer's even-handed treatment of various site types reflects a balanced approach that respects different visitor interests. *Step 4:* The Loyalty/Betrayal foundation is mildly engaged through the appreciation for local cuisine and traditions, honoring the regional identity of Piedmont. *Step 5:* The neutral, culturally appreciative tone fits the persona's left-leaning orientation by celebrating cultural diversity and respect for heritage without being preachy or ideological.

Population Aligned Language Models RATIONALE — REJECTED

Personality. *Step 1:* The bullet-point list format is clear and organized, which appeals to conscientiousness, but the briefer descriptions may feel less engaging for a young persona with high openness who likely wants vivid storytelling. *Step 2:* The simpler structure is functional but lacks the descriptive depth that often appeals to a culturally curious graduate looking for meaningful travel guidance. *Step 3:* The list misses sensory or culinary highlights that would engage extraversion through the social pleasures of food and dining, which often matter to young travelers. *Step 4:* The agreeable side is engaged through the inviting tone, but the formal list structure feels less warm than a flowing narrative would feel for the persona. *Step 5:* The straightforward format keeps things simple, suiting low neuroticism, but the lighter detail may leave the persona wanting more atmosphere and texture in the recommendations.

Cultural Dimensions. *Step 1:* The list format reflects strong uncertainty avoidance, which is helpful for planning, but may feel slightly less aligned with the more flowing cultural narrative that often resonates with Brazilian sensibilities. *Step 2:* The absence of local cuisine recommendations underplays the indulgence-oriented framing that suits both Brazilian culture and a young traveler's enjoyment of food experiences. *Step 3:* The brief, structured descriptions feel slightly more masculine in cultural orientation, focusing on efficient information delivery rather than the feminine warmth of sensory cultural immersion. *Step 4:* The long-term orientation toward cultural appreciation is present but less developed without descriptive richness about why each site matters. *Step 5:* The cultural depth feels slightly thinner overall, which may be less satisfying for a culturally engaged young traveler seeking immersive experiences.

Human Values. *Step 1:* The Stimulation value is supported through the variety of attractions, but the brief descriptions reduce the excitement and richness of each option. *Step 2:* The Hedonism value is weakly engaged because the answer omits cuisine and sensory pleasures that often bring travel to life. *Step 3:* The Universalism value is touched on through the cultural variety, but the lighter treatment doesn't as strongly engage the persona's likely left-leaning appreciation for cultural depth. *Step 4:* The Self-Direction value is partially supported through the list format that lets the persona pick what interests him, but the lack of detail provides less basis for informed choice. *Step 5:* The Tradition value is engaged through historical sites, but without the cultural color of cuisine and atmosphere, the heritage appreciation feels less rich.

Primal World Beliefs. *Step 1:* The Enticing world belief is engaged through the variety of attractions, but the brief descriptions reduce the sense of wonder and excitement that a more vivid response would convey. *Step 2:* The Abundant world belief is supported through the list of multiple sites, similar to the alternative, but feels less rich without descriptive context. *Step 3:* The Alive world belief is weakly engaged because the list format treats sites more as information points than as living parts of the city's culture. *Step 4:* The Interconnected world belief is partially supported through the variety of cultural elements, but without flowing description, the connections between sites feel less developed. *Step 5:* The Safe world belief is implicitly engaged through the orderly format, but the answer feels less inviting than a more narrative approach.

Moral Foundations. *Step 1:* The Care/Harm foundation is mildly engaged through the appreciative tone, but the briefer treatment feels less warm than a more descriptive approach. *Step 2:* The Sanctity/Degradation foundation is touched on through references to historical sites, but with lighter framing about their significance. *Step 3:* The Fairness/Cheating foundation is not particularly relevant here, but the balanced list of site types reflects fair treatment of different visitor interests. *Step 4:* The Loyalty/Betrayal foundation is weakly engaged because the answer omits local cuisine and traditions that would honor regional identity more strongly. *Step 5:* The neutral tone fits the persona's left-leaning orientation, but the moral framing of cultural appreciation feels lighter and less integrated than a richer treatment.

Table 19: **Test-time rationale examples for India.**

Persona. You are an 18–34-year-old male from India with Indo-Aryan background, post-secondary graduate (some or complete graduate degree), and a centrist political orientation.

Question. I want to improve my leadership skills and set specific goals for professional development.

Chosen Answer. To improve your leadership skills and set specific goals for professional development, the following steps can be taken: (1) Self-assessment: evaluate your current strengths and weaknesses to identify areas needing improvement. (2) Set specific goals: define SMART (Specific, Measurable, Achievable, Relevant, Time-bound) goals—e.g., "improve team communication by 20% within the next 6 months." (3) Participate in leadership development programs: join training, workshops, or online courses. (4) Find a mentor: learn from experienced leaders who can guide you and evaluate your progress. (5) Seek regular feedback: get input from your team and colleagues. (6) Maintain flexibility: adapt your strategy based on new challenges.

Rejected Answer. To improve leadership skills and set specific goals for professional development, evaluate your current situation and identify your weaknesses. Then set SMART (Specific, Measurable, Achievable, Relevant, Time-bound) goals—for example, aim to improve communication with your team over the next 6 months or successfully complete a specific project. Regular self-assessment and feedback are also important to track progress and adjust your strategy as needed.

GENERIC-*Population Aligned Language Models* RATIONALE — CHOSEN

Step 1: The answer provides a clear, well-structured set of steps for leadership development, which is likely to appeal to a young graduate seeking organized career guidance. *Step 2:* The variety of recommendations—self-assessment, SMART goals, programs, mentorship, feedback—gives the persona multiple paths for growth that he can adapt to his situation. *Step 3:* The mentorship and team feedback elements align with the persona's likely cultural preference for learning through community and respected figures. *Step 4:* The neutral, professional tone fits the persona's centrist political orientation by avoiding any ideological framing and staying focused on practical career development. *Step 5:* Overall, the answer balances structure, variety, and relational elements, making it a strong fit for a young Indian graduate looking to grow as a leader.

GENERIC-*Population Aligned Language Models* RATIONALE — REJECTED

Step 1: The answer covers the basics of leadership development but with less detail and variety, which may feel limiting for a young graduate seeking comprehensive career guidance. *Step 2:* The lack of mentorship and program recommendations underplays the cultural value placed on learning through community and experienced figures in the persona's Indian context. *Step 3:* The narrower set of suggestions makes the path to growth feel less rich, which may not fully energize a young professional with multiple ambitions. *Step 4:* The neutral tone fits the persona's centrist orientation, but the briefer framing offers less actionable depth for someone serious about building leadership skills. *Step 5:* Overall, while the answer is clear and includes the core SMART goals concept, its limited scope makes it a slightly weaker fit for a young Indian graduate looking for a fuller, more well-rounded leadership development plan.

Population Aligned Language Models RATIONALE — CHOSEN

Personality. *Step 1:* The persona is a young Indian male with a graduate degree, which suggests high openness to experience and a willingness to learn new skills. The answer's variety of steps—self-assessment, mentorship, programs, feedback—appeals to his curiosity and desire to grow. *Step 2:* The structured, numbered list of steps reflects strong conscientiousness, giving the persona a clear plan he can follow. As a young professional building his career, he likely values this kind of organized approach. *Step 3:* The recommendation to find a mentor and seek feedback taps into a balanced extraversion profile, encouraging social engagement and learning from others. The persona may find this energizing and useful for his career growth. *Step 4:* The agreeable tone of seeking feedback and adapting based on others' input fits well with the persona's likely cooperative side, showing that leadership involves collaboration rather than just personal achievement. *Step 5:* The reminder to stay flexible reduces pressure and supports a persona with low to moderate neuroticism who appreciates adaptable strategies rather than rigid plans that could cause stress when things change.

Cultural Dimensions. *Step 1:* The recommendation to find a mentor reflects a moderate power distance acceptance, which fits Indian cultural norms that value learning from experienced and respected figures. The persona likely finds this familiar and useful in his career context. *Step 2:* The emphasis on community-based learning—through mentors, programs, and team feedback—aligns with collectivist cultural elements common in Indian society, where personal growth often happens within social networks rather than alone. *Step 3:* The structured SMART goals framework supports a moderate uncertainty avoidance preference, giving the persona predictable methods for tracking progress. As a graduate, he likely appreciates this kind of organized approach. *Step 4:* The focus on long-term skill development and steady improvement aligns with a long-term orientation, which fits both Indian cultural values of perseverance and the persona's likely career-building mindset. *Step 5:* The balance between achievement-focused goal-setting and relational elements like mentorship and feedback reflects a moderate masculine-feminine cultural blend, giving the persona room to pursue success while also valuing relationships.

Human Values. *Step 1:* The Achievement value is strongly supported through the focus on goal-setting, skill-building, and measurable progress, which fits the persona's likely drive to demonstrate competence in his career. *Step 2:* The Self-Direction value is engaged through the emphasis on self-assessment and personal goal-setting, giving the persona room to define his own path while still being structured. *Step 3:* The Benevolence value is supported through the mentorship and feedback recommendations, building positive relationships with others in his professional network. *Step 4:* The Security value is reflected in the structured, step-by-step framework, giving the persona confidence that he is following a reliable approach to skill development. *Step 5:* The recommendation to stay flexible touches on the Stimulation value, encouraging the persona to remain open to new challenges and adapt as his career evolves.

continued on next page

Table 19 continued.

Primal World Beliefs. *Step 1:* The answer presents the world of professional development as Abundant, with many resources—programs, mentors, feedback opportunities—available to support the persona's growth. *Step 2:* The Just world belief is supported because the answer suggests that consistent effort, self-assessment, and feedback lead to real improvement, reinforcing the idea that hard work pays off. *Step 3:* The Interconnected world belief is engaged through the emphasis on learning from mentors and team feedback, showing that growth happens within a network of relationships rather than in isolation. *Step 4:* The Enticing framing of leadership development—full of programs, workshops, and learning opportunities—presents the professional world as full of possibility for a young graduate. *Step 5:* The world is presented as Safe enough to experiment with new strategies, with flexibility built in to handle challenges. This supports the persona's confidence in pursuing growth.

Moral Foundations. *Step 1:* The Authority/Subversion foundation is engaged through the recommendation to find a mentor, honoring the legitimate authority of experienced leaders. The persona likely respects this guidance pattern from his cultural and educational background. *Step 2:* The Fairness/Cheating foundation is supported through the focus on regular feedback, ensuring that progress is honestly evaluated rather than just self-claimed. *Step 3:* The Care/Harm foundation is engaged through the emphasis on building relationships and seeking input from team members, showing care for those the persona will lead. *Step 4:* The Loyalty/Betrayal foundation is touched on through the team-oriented framing, encouraging commitment to one's colleagues and shared goals. *Step 5:* The overall tone is balanced and non-ideological, fitting the persona's centrist political orientation by focusing on practical professional growth rather than any partisan framing.

Population Aligned Language Models RATIONALE — REJECTED

Personality. *Step 1:* The answer covers the basics but with less variety, which may feel limiting for a young persona with high openness who likely wants to explore multiple paths to leadership growth. *Step 2:* The conscientiousness side is partially served by the SMART goals framework, but the lack of clear step-by-step structure may feel less organized for a graduate who values methodical guidance. *Step 3:* The answer's brief mention of self-assessment and feedback engages extraversion only weakly, missing the social and mentorship dimensions that often energize young professionals building their careers. *Step 4:* The agreeable tone is present through the feedback recommendation, but without mentorship or team-oriented suggestions, the answer feels less collaborative than a young leader might want. *Step 5:* The shorter framing keeps things simple but may feel slightly under-developed for a persona with low neuroticism who likely wants a comprehensive plan he can rely on for career growth.

Cultural Dimensions. *Step 1:* The lack of mentorship recommendation underplays the moderate power distance acceptance common in Indian culture, where learning from experienced figures is highly valued. *Step 2:* The absence of community-based learning suggestions like training programs or workshops feels less aligned with collectivist elements of Indian cultural traditions, which often emphasize learning in groups. *Step 3:* The SMART goals framework does provide moderate uncertainty avoidance, but without additional structured steps, the answer offers less predictability than a more detailed approach. *Step 4:* The long-term orientation is partially supported through the goal-setting framework, but the briefer treatment doesn't as strongly reinforce the patient, steady cultivation of leadership that the persona likely values. *Step 5:* The answer leans more toward individual achievement without strongly balancing it with relational and community-focused elements, making the cultural fit slightly weaker for an Indian graduate.

Human Values. *Step 1:* The Achievement value is supported through goal-setting, but the narrower set of recommendations underweights the breadth of competence-building options the persona likely seeks. *Step 2:* The Self-Direction value is engaged through self-assessment, but the answer doesn't fully empower the persona with multiple paths to choose from for his independent growth. *Step 3:* The Benevolence value is weakly engaged because the answer doesn't include mentorship or team-building elements, missing the relational growth dimension. *Step 4:* The Security value is partially supported through the SMART framework, but the simpler structure may feel less reassuring to a persona who values comprehensive, reliable career guidance. *Step 5:* The Stimulation value is mildly engaged through the mention of adjusting strategy, but the answer doesn't offer as much variety or excitement in its development paths as a richer treatment would.

Primal World Beliefs. *Step 1:* The Abundant world belief is mildly underrepresented because the answer focuses on a few core practices rather than highlighting the wide range of resources available for professional growth. *Step 2:* The Just world belief is supported through the SMART goals and feedback framing, showing that effort yields visible results, which the persona likely appreciates. *Step 3:* The Interconnected world belief is weakly engaged because the answer focuses on individual self-assessment and goal-setting rather than learning through mentors, peers, or community. *Step 4:* The Enticing framing is present but limited, since the answer doesn't paint as full a picture of the opportunities available in leadership development as a more detailed response would. *Step 5:* The world is implicitly Safe in this approach, but the narrower scope may feel less expansive than the broader landscape of growth options the persona likely envisions.

Moral Foundations. *Step 1:* The Authority/Subversion foundation is weakly engaged because the answer doesn't include mentorship or learning from experienced leaders, missing a key cultural and moral element. *Step 2:* The Fairness/Cheating foundation is supported through the feedback element, but the briefer treatment doesn't as strongly emphasize honest evaluation as a richer response would. *Step 3:* The Care/Harm foundation is partially engaged through team communication goals, but without explicit relational suggestions, the moral dimension of caring for one's team feels lighter. *Step 4:* The Loyalty/Betrayal foundation is barely touched, since the answer doesn't strongly emphasize commitment to one's team or community. *Step 5:* The overall tone is professional and centrist, which fits the persona, but the moral framing feels less integrated and complete compared to a fuller treatment of leadership ethics.

Table 20: **Test-time rationale examples for France.**

Persona. You are a 35–45-year-old male from France with immigrant background, centrist political orientation, and some post-secondary education.

Question. I own a small retail shop and need help writing a social media post to announce a new campaign, including special promotions and limited-time offers.

Chosen Answer. Focus on highlighting the community aspect of your shop by emphasizing how these promotions and limited-time offers contribute to the well-being of the local community. You can mention partnerships with local brands, initiatives supporting local artisans, or social responsibility actions your shop undertakes. This approach helps strengthen the emotional connection with your customers and promotes a positive image of your business.

Rejected Answer. Create a post that highlights the exceptional savings your customers can enjoy through special promotions and limited-time offers. You can include details about products on sale, discount percentages, and deadlines for the offers. Don't forget to incorporate attractive visual elements such as images or videos of the products to capture your followers' attention.

GENERIC-*Population Aligned Language Models* RATIONALE — CHOSEN

Step 1: The answer focuses on community connection and local partnerships, which is likely to appeal to a small shop owner who depends on neighborhood relationships for his business success. *Step 2:* The recommendations around supporting local artisans and social responsibility align with French cultural values around community and quality, fitting the persona's setting. *Step 3:* The persona's immigrant background may make him especially appreciative of community-focused framing that builds belonging and emotional ties with customers. *Step 4:* The neutral, ethical tone respects the persona's centrist political orientation by focusing on universally appreciated values like community and care without ideological framing. *Step 5:* Overall, the answer balances brand-building, community focus, and ethical business practices, making it a strong fit for a middle-aged French shop owner planning a meaningful campaign.

GENERIC-*Population Aligned Language Models* RATIONALE — REJECTED

Step 1: The answer focuses on savings and visual appeal, which is practical but may feel less meaningful for a small shop owner who likely values community connection and authentic brand-building. *Step 2:* The transactional framing underweights the relational and community-focused values that often resonate with French cultural traditions and immigrant business experiences. *Step 3:* The recommendations around deadlines and discount percentages are useful but feel less complete than a strategy that includes local partnerships and social responsibility. *Step 4:* The neutral tone fits the persona's centrist orientation, but the narrow promotional focus offers less depth for someone building a long-term community presence. *Step 5:* Overall, while the answer is clear and actionable, its lighter community focus makes it a slightly weaker fit for a middle-aged French shop owner with immigrant roots seeking a meaningful campaign approach.

Population Aligned Language Models RATIONALE — CHOSEN

Personality. *Step 1:* The persona is a middle-aged French male shop owner with an immigrant background, which suggests he values community connection and authenticity. The answer's emphasis on local community ties resonates well with his likely agreeable nature that prioritizes meaningful relationships with customers. *Step 2:* The recommendation to highlight partnerships with local brands and artisans appeals to openness to experience, encouraging the persona to think creatively about how his shop fits into the broader local ecosystem. *Step 3:* The conscientious side is supported through the structured approach of building a positive brand image step by step, which helps the persona plan a thoughtful marketing strategy. *Step 4:* The community-focused framing engages a moderate extraversion profile, encouraging social and relational engagement with customers and local partners, which suits a small business owner who likely thrives on community connections. *Step 5:* The answer reduces pressure by focusing on relationship-building rather than aggressive sales tactics, fitting a persona with low to moderate neuroticism who likely prefers sustainable and ethical marketing strategies over pushy promotions.

Cultural Dimensions. *Step 1:* The emphasis on local community and supporting local artisans reflects a collectivist cultural orientation, which fits well with French cultural values around community, terroir, and supporting local craftsmanship. *Step 2:* The persona's immigrant background may give him a particularly strong appreciation for community connection, which the answer honors through its focus on emotional ties between business and neighborhood. *Step 3:* The framing of social responsibility actions aligns with feminine cultural dimensions valuing care and quality of life over pure competition, which fits French and European cultural sensibilities around ethical business. *Step 4:* The long-term orientation is reflected in building lasting customer relationships through community engagement rather than one-time discount-driven transactions, supporting sustained brand loyalty. *Step 5:* The moderate uncertainty avoidance is engaged through the steady, relationship-focused approach that gives the persona a reliable strategy for building a strong reputation in his neighborhood.

Human Values. *Step 1:* The Benevolence value is strongly supported through the focus on community wellbeing and supporting local artisans, which fits the persona's likely care for those he sees in his daily business life. *Step 2:* The Universalism value is engaged through the broader concern for community welfare and social responsibility, reflecting a generally caring approach to business that respects all stakeholders. *Step 3:* The Tradition value is partially supported through the appreciation for local artisans and craftsmanship, honoring the cultural heritage that gives a small business its character. *Step 4:* The Self-Direction value is engaged through the encouragement for the persona to define his own brand story rooted in community, giving him room to express his unique business identity. *Step 5:* The Security value is reflected in the long-term reputation-building approach, helping the persona create a stable and trusted business presence in his neighborhood.

continued on next page

Table 20 continued.

Primal World Beliefs. *Step 1:* The answer presents the world as Interconnected, where the shop, customers, local artisans, and broader community are all linked in meaningful ways. The persona likely sees his business this way, embedded in a web of relationships. *Step 2:* The Abundant world belief is supported through the recognition that there are many local partners and community resources to draw upon, reinforcing a sense of plenty within the local ecosystem. *Step 3:* The Alive world belief is engaged through the framing of customer connections as emotional and meaningful rather than purely transactional, suggesting that business interactions carry real human weight. *Step 4:* The Safe world belief is touched on through the focus on building trust and a positive image, presenting the local community as a reliable and supportive context for the business. *Step 5:* The Just world belief is mildly engaged through the idea that ethical, community-focused business behavior leads to a positive reputation and customer loyalty.

Moral Foundations. *Step 1:* The Care/Harm foundation is strongly engaged through the focus on community wellbeing and supporting local artisans, expressing genuine concern for the people around the business. *Step 2:* The Loyalty/Betrayal foundation is supported through the emphasis on emotional connection with customers and partnerships with local brands, building a sense of mutual commitment within the neighborhood. *Step 3:* The Fairness/Cheating foundation is mildly engaged through the social responsibility framing, suggesting ethical business practices that treat all stakeholders fairly. *Step 4:* The Authority/Subversion foundation is weakly but positively engaged through the respect for local craftsmanship and traditional artisan practices that anchor the community. *Step 5:* The neutral tone fits the persona's centrist orientation by avoiding polarizing framing while still engaging meaningful values around community and ethical business conduct.

Population Aligned Language Models RATIONALE — REJECTED

Personality. *Step 1:* The answer focuses on savings and discounts, which is practical but may feel less meaningful for a persona who likely values authentic community connection in his business. *Step 2:* The conscientious side is partially served through specific advice about deadlines and discount percentages, but the narrow focus on transactional details may feel less complete than a broader strategy. *Step 3:* The recommendation to use attractive visuals appeals to openness, but the overall framing leans toward a generic promotional approach that lacks the relational depth a small shop owner often values. *Step 4:* The agreeable side is weakly engaged because the answer focuses on extracting customer attention through deals rather than building genuine emotional connection with the local community. *Step 5:* The aggressive promotional framing may feel slightly stress-inducing for a persona with low neuroticism who likely prefers steady, relationship-based business growth over urgent sales tactics.

Cultural Dimensions. *Step 1:* The focus on individual savings and discounts leans toward an individualist cultural orientation, which may feel slightly less aligned with the collectivist community-focused values often important in French culture and to immigrant business owners. *Step 2:* The transactional approach underplays the cultural appreciation for craftsmanship and local heritage that often shapes how French shops build their identities. *Step 3:* The high-energy promotional framing reflects a more masculine cultural orientation focused on competition and visible success, which may feel less aligned with the persona's likely preference for steady, quality-focused business. *Step 4:* The short-term orientation of limited-time offers and quick discounts feels less aligned with a long-term orientation toward building lasting customer relationships and brand loyalty. *Step 5:* The lack of community framing reduces the cultural fit for an immigrant shop owner who likely sees his business as embedded in a meaningful neighborhood context.

Human Values. *Step 1:* The Achievement value is supported through the focus on attracting customers and driving sales, but this narrower business framing underweights the broader values the persona likely brings to his work. *Step 2:* The Benevolence value is weakly engaged because the answer doesn't strongly address customer or community wellbeing beyond extracting their attention with deals. *Step 3:* The Universalism value is barely touched, since the answer doesn't engage with social responsibility or broader community welfare. *Step 4:* The Hedonism value is mildly supported through the appeal of savings and exciting deals, but this is a thin engagement compared to richer relational values. *Step 5:* Overall, the answer leans heavily on Achievement and Hedonism, leaving Benevolence, Universalism, and Tradition underrepresented—an imbalance that may feel less aligned with the persona's likely value mix.

Primal World Beliefs. *Step 1:* The Abundant world belief is engaged through the focus on plentiful deals and savings, but this is a more transactional form of abundance that lacks the relational richness of community connection. *Step 2:* The Just world belief is mildly supported through the implicit promise that good deals attract loyal customers, but this feels somewhat shallow compared to deeper trust-building. *Step 3:* The Interconnected world belief is weakly engaged because the answer focuses on the shop's individual promotions rather than its embeddedness in the local community. *Step 4:* The Enticing framing comes through in the visual appeal and deal-driven excitement, but it leans more toward stimulation than meaningful connection. *Step 5:* The world is presented as Safe enough through the structured promotional approach, but the absence of relational framing reduces the depth of the worldview the answer supports.

Moral Foundations. *Step 1:* The Care/Harm foundation is weakly engaged because the answer focuses on capturing attention through deals rather than expressing care for customers and community. *Step 2:* The Loyalty/Betrayal foundation is touched on through customer attraction, but without community framing, the moral commitment to neighborhood ties is largely absent. *Step 3:* The Fairness/Cheating foundation is mildly supported through clear discount transparency, but this feels limited compared to broader fairness in business practices. *Step 4:* The Authority/Subversion foundation is barely engaged, since the answer doesn't address respect for local craftsmanship or community traditions. *Step 5:* Overall, the moral framing leans heavily on transactional appeal, which is valid but less aligned with the relational, community-respecting moral orientation the persona likely brings to his work.

Table 21: **Test-time rationale examples for Italy.**

Persona. You are an 18–34-year-old male from Italy with Indo-European background, post-secondary graduate (some or complete graduate degree), and a centrist political orientation.

Question. I am attending a baby naming party for a friend's child and need help choosing a gift and writing a sincere greeting card that expresses congratulations and wishes for the new arrival in the family.

Chosen Answer. A practical gift such as a set of baby clothes or toys would be greatly appreciated. For the greeting card, you could write: "I am so happy for you and your family! The arrival of a baby is an incredible moment, and I hope you can enjoy every part of this journey. My wish is that your child is blessed with a life full of laughter, adventures, and success, and that your family is always supported and loved."

Rejected Answer. A personalized gift such as a photo album or a decorative item for the baby's nursery can be a great choice. For the greeting card, you could write: "Congratulations on the arrival of your little treasure! I hope this new chapter of your life is filled with joy, love, and serenity. My wish is that your child grows up healthy, happy, and curious, and that your family is always united and full of love."

GENERIC-*Population Aligned Language Models* RATIONALE — CHOSEN

Step 1: The answer provides a practical gift suggestion and a warm, sincere greeting card, which is likely to appeal to a young graduate looking for thoughtful but uncomplicated guidance. *Step 2:* The recommendation of baby clothes or toys offers something useful that the friend's family will actually appreciate, fitting the persona's practical mindset. *Step 3:* The greeting card focuses on family support and love, aligning with Italian cultural values around close family bonds. *Step 4:* The neutral, heartfelt tone respects the persona's centrist orientation by avoiding any ideological framing while still feeling sincere. *Step 5:* Overall, the answer balances practicality and warmth, making it a strong fit for a young Italian graduate attending a baby naming party for a close friend.

GENERIC-*Population Aligned Language Models* RATIONALE — REJECTED

Step 1: The answer provides a thoughtful greeting card and a personalized gift idea, but the decorative item or photo album may feel less practical than something the family can use day-to-day. *Step 2:* The card's focus on serenity and curiosity is sweet but may feel slightly more abstract than wishes tied to family support and tangible joy. *Step 3:* The personalized gift adds an element of taste-matching that may feel less reliable than a clearly useful gift for a young graduate who likely prefers straightforward choices. *Step 4:* The neutral tone fits the persona's centrist orientation, but the slightly more poetic framing may feel less natural than a simpler, warmer expression. *Step 5:* Overall, while the answer is sincere and warm, its more abstract gift and card style make it a slightly weaker fit for a young Italian graduate seeking practical and heartfelt guidance for a friend's baby celebration.

Population Aligned Language Models RATIONALE — CHOSEN

Personality. *Step 1:* The persona is a young Italian male graduate, which suggests practical thinking and an appreciation for useful, well-thought-out gifts. The recommendation of baby clothes or toys aligns with his conscientious side that values gifts which serve real needs. *Step 2:* The straightforward gift suggestion appeals to a balanced openness profile—open enough to consider thoughtful options, but practical enough to choose something the family will actually use. The persona likely finds this groundedness appealing. *Step 3:* The warm and friendly tone of the greeting card fits well with a moderate extraversion profile, expressing happiness and connection without being overly dramatic. The persona may appreciate this natural, sincere voice. *Step 4:* The mention of wishingthe family to be "always supported and loved" reflects high agreeableness, focusing on care for the family unit rather than just the child. The persona, attending a friend's celebration, likely values this relational warmth. *Step 5:* The answer keeps things simple and warm, suiting a persona with low to moderate neuroticism who likely doesn't want to overthink the gift-giving process and prefers easy, sincere expressions of support.

Cultural Dimensions. *Step 1:* The recommendation of practical baby clothes or toys reflects an appreciation for usefulness, which fits a balanced cultural approach common in Italian society where thoughtful practicality is valued in family gifts. *Step 2:* The greeting card's emphasis on family support and love aligns with collectivist cultural elements common in Italian culture, where family bonds and intergenerational support are deeply valued. *Step 3:* The straightforward tone of the gift recommendation reflects moderate uncertainty avoidance, giving the persona a clear and reliable suggestion he can act on without overthinking. *Step 4:* The wish for a life of "laughter, adventures, and success" balances feminine cultural values like joy and warmth with masculine values like achievement, fitting a young Italian graduate's likely blended outlook. *Step 5:* The long-term orientation comes through in the wish for ongoing family support and a successful life journey, reflecting a future-focused perspective that resonates with the persona's likely stage in life.

Human Values. *Step 1:* The Benevolence value is strongly supported through the focus on family wellbeing and lasting support, which fits the persona's likely care for his friend's family during this meaningful moment. *Step 2:* The Achievement value is touched on through the wish for the child's success, acknowledging the persona's likely respect for personal accomplishment as a graduate himself. *Step 3:* The Hedonism value is engaged through the wish for laughter and joyful adventures, validating the celebration of life's pleasures during this happy occasion. *Step 4:* The Security value is reflected in the wish for the family to be "always supported and loved," offering a sense of stability that the persona likely values for his friend's growing family. *Step 5:* The Universalism value is mildly engaged through the broad, generous wishes for the child's life, reflecting a generally caring outlook that fits the persona's centrist orientation.

continued on next page

Table 21 continued.

Primal World Beliefs. *Step 1:* The answer presents the world as Enticing and full of opportunity for the new child, with wishes for adventures and a life full of laughter. The persona likely sees the world this way for a friend's newborn. *Step 2:* The Abundant world belief is supported through the generous wishes for laughter, adventures, and success, suggesting that life has plenty to offer the new arrival. *Step 3:* The Interconnected world belief is engaged through the focus on family support and love, framing the child's life as embedded in a network of caring relationships. *Step 4:* The Safe world belief comes through in the wish for the family to be "always supported and loved," presenting the world as a place where family bonds provide reliable security. *Step 5:* The Just world belief is implicitly supported through the wish that the child enjoys success, reflecting a sense that effort and goodness are rewarded in life.
Moral Foundations. *Step 1:* The Care/Harm foundation is strongly engaged through the warm wishes for the child and the family, expressing genuine concern for their wellbeing. The persona likely shares this caring orientation toward his friend. *Step 2:* The Loyalty/Betrayal foundation is supported through the focus on family unity and ongoing support, honoring the importance of close bonds. *Step 3:* The Fairness/Cheating foundation is mildly engaged through the wish for success, suggesting that the child's efforts will be justly rewarded throughout life. *Step 4:* The Sanctity/Degradation foundation is touched on in the framing of birth as "an incredible moment," treating the arrival of a new life with appropriate reverence. *Step 5:* The neutral tone respects the persona's centrist orientation by avoiding any ideological framing and focusing on universally appreciated values like family, love, and joy.

Population Aligned Language Models RATIONALE — REJECTED
Personality. *Step 1:* The suggestion of a personalized photo album or decorative item appeals to openness, but it may feel less practical for a young graduate who likely prefers gifts the family will use day-to-day. *Step 2:* The conscientious side may find the decorative gift suggestion less functional than baby clothes or toys, since decorative items don't serve a clear daily need for the family. *Step 3:* The greeting card's tone is warm and sincere, fitting a balanced extraversion profile, but the framing leans more toward poetic expression than the natural, conversational warmth a young persona might prefer. *Step 4:* The agreeable tone is present through the focus on family unity, but the slightly more flowery language may feel less genuine compared to a simpler, more direct expression of support. *Step 5:* The decorative gift idea may add slight overthinking for a persona with low neuroticism who likely prefers easy, useful gift choices over ones that require more curation and personal taste judgment.
Cultural Dimensions. *Step 1:* The personalized photo album or nursery decoration suggestion reflects an individualist cultural framing focused on personal aesthetic, which may feel slightly less aligned with the practical family-focused orientation common in Italian gift-giving traditions. *Step 2:* The greeting card emphasizes serenity and curiosity, which is lovely but may feel slightly less aligned with the warmth and joyful expressiveness often valued in Italian family celebrations. *Step 3:* The decorative gift suggestion introduces moderate uncertainty since it depends on matching the family's taste, which may feel less reliable than a clearly useful gift like clothes or toys. *Step 4:* The card leans toward a more feminine cultural orientation with its focus on joy, love, and serenity, which is fine but may not as strongly balance the achievement-oriented values the persona may also appreciate. *Step 5:* The long-term orientation is supported through wishes for the child's growth, but the slightly more abstract framing feels less grounded than wishes tied to concrete family support.
Human Values. *Step 1:* The Benevolence value is supported through the warm wishes for the family, but the gift choice of decorative items may feel less directly caring than something practical the family can use. *Step 2:* The Achievement value is more weakly engaged because the card focuses on serenity and happiness rather than success, which a young graduate may also value. *Step 3:* The Hedonism value is touched on through wishes for joy and love, but the more contemplative framing feels less celebratory than a more lively expression. *Step 4:* The Security value is partially supported through wishes for family unity, but the decorative gift adds an element of uncertainty about whether the family will love the choice. *Step 5:* The Universalism value is mildly engaged through the broad caring wishes, but the abstract decorative gift suggestion makes the overall package feel slightly less grounded in tangible care.
Primal World Beliefs. *Step 1:* The Enticing world belief is engaged through wishes for joy and curiosity, but the framing feels slightly more contemplative than fully celebratory of life's possibilities. *Step 2:* The Abundant world belief is partially supported through the wishes for the child's growth, but the decorative gift choice limits the sense of practical generosity. *Step 3:* The Interconnected world belief is engaged through the focus on family unity, similar to the alternative, so this belief is reasonably well-supported. *Step 4:* The Safe world belief is touched on through wishes for serenity, but the decorative gift adds a small layer of uncertainty about whether it will fit the family's space. *Step 5:* The Just world belief is less directly engaged because the card focuses more on emotional states than on the rewarded effort and success that often resonate with a young graduate.
Moral Foundations. *Step 1:* The Care/Harm foundation is engaged through warm wishes for the child and family, similar to the alternative, so this moral element is reasonably well-supported. *Step 2:* The Loyalty/Betrayal foundation is touched on through the focus on family unity, but the framing feels slightly less grounded in ongoing tangible support. *Step 3:* The Fairness/Cheating foundation is weakly engaged because the card focuses on emotional states rather than success or earned outcomes. *Step 4:* The Sanctity/Degradation foundation is gently engaged through references to a "little treasure," treating the child's arrival as precious, though the framing feels slightly more sentimental than reverent. *Step 5:* The neutral tone fits the persona's centrist orientation, but the moral framing feels slightly less integrated and complete than a more grounded treatment of family values.